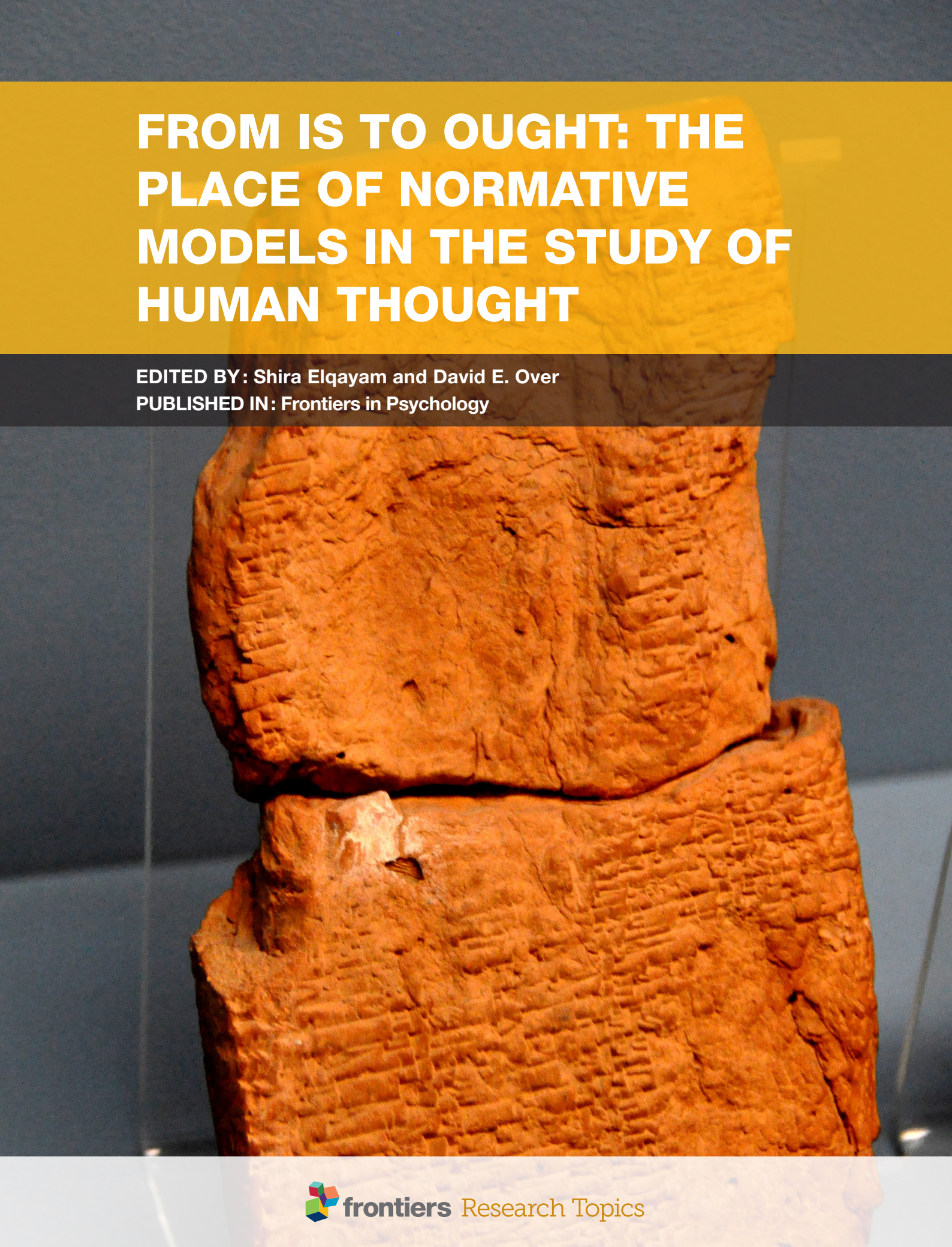
FROM IS TO OUGHT: THE PLACE OF NORMATIVE MODELS IN THE STUDY OF HUMAN THOUGHT

EDITED BY: Shira Elqayam and David E. Over
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-896-2
DOI 10.3389/978-2-88919-896-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

FROM IS TO OUGHT: THE PLACE OF NORMATIVE MODELS IN THE STUDY OF HUMAN THOUGHT

Topic Editors:
Shira Elqayam, De Montfort University, UK
David E. Over, Durham University, UK

Written in Sumerian in Mesopotamia, (circa 2060-2050 BC), the Law Code of King Ur-Nammu is the oldest known normative code that survives today. This tablet is on display at the Istanbul Archaeological Museum, Turkey.

In the study of human thinking, two main research questions can be asked:

> "Descriptive Q: What *is* human thinking like?
> Normative Q: What *ought* human thinking be like?"

For decades, these two questions have dominated the field, and the relationship between them generated many a controversy. *Empirical normativist* approaches regard the answers to these questions as *positively correlated* – in essence, human thinking is what it ought to be (although

what counts as the 'ought' standard is moot). In contemporary theories of reasoning and decision making, this is often associated with a Panglossian framework, an adaptationist approach which regards human thinking as *a priori* rational.

In contrast, *prescriptive normativism* sees the answers to these two questions as *negatively correlated*. Normative models are still relevant to human thought, but human behaviour deviates from them quite markedly (with the invited conclusion that humans are often irrational). Prescriptive normativism often results in a Meliorist agenda, which sees rationality as amenable to education.

Both empirical and prescriptive normativism can be contrasted with a *descriptivist* framework for psychology of human thinking. Following Hume's strict divide between the 'is' and the 'ought', descriptivism regards the descriptive and normative research questions as *uncorrelated*, or dissociated, with only the former question suitable for psychological study of human behaviour.

This basic division carries over to the relation between normative ('ought') rationality, based on conforming to normative standards; and instrumental ('is') rationality, based on achieving one's goals. Descriptivist approaches regard the two as dissociated, whereas normativist approaches tend to see them as closely linked, with normative arguments defining and justifying instrumental rationality.

This research topic brings together diverse contributions to the continuing debate. Featuring contributions from leading researchers in the field, the e-book covers a wide range of subjects, arranged by six sections:

The standard picture: Normativist perspectives
In defence of soft normativism
Exploring normative models
Descriptivist perspectives
Evolutionary and ecological accounts
Empirical reports

With a total of some 24 articles from 55 authors, this comprehensive treatment includes theoretical analyses, meta-theoretical critiques, commentaries, and a range of empirical reports. The contents of the Research Topic should appeal to psychologists, linguists, philosophers and cognitive scientists, with research interests in a wide range of domains, from language, through reasoning, judgment and decision making, and moral judgment, to epistemology and theory of mind, philosophical logic, and meta-ethics.

Citation: Elqayam, S., Over, D. E., eds. (2016). From Is to Ought: The Place of Normative Models in the Study of Human Thought. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-896-2

Table of Contents

Evolutionary and ecological accounts

Empirical reports

frontiers in Psychology

EDITORIAL
published: 28 April 2016
doi: 10.3389/fpsyg.2016.00628

Editorial: From Is to Ought: The Place of Normative Models in the Study of Human Thought

Shira Elqayam[1]* *and David E. Over*[2]

[1] School of Applied Social Sciences, De Montfort University, Leicester, UK, [2] Department of Psychology, Durham University, Durham, UK

Keywords: is-ought inference, meliorism, Panglossianism, descriptisim, rationality debate, normativism

The Editorial on the Research Topic

From Is to Ought: The Place of Normative Models in the Study of Human Thought

Normative rules and regulations are everywhere we turn; they are, as Searle (2005) memorably called them, the glue that holds human society together. We stop at red lights and try (not always very successfully) to be fair and truthful in our personal and professional lives. We humans are the only species that internalizes normative rules (Carruthers, 2006), and feels shame and guilt when we violate them. Moreover, we humans are the only species capable of creating novel norms from scratch (Elqayam et al., 2015)—a species-specific, generative capacity no less extraordinary than the much-celebrated generative capacity to create novel sentences. It is not surprising, then, that normative rules feature so prominently in much of the psychology of higher mental processing—reasoning, decision making, and moral judgment. Normative rules dominate much of the great rationality debate, mainly in the form of the striking *normative-descriptive gap* (Stanovich and West, 2000). Human behavior often deviates from formal standards of rationality, such as classical logic and probability theory.

Can humans be said to be rational at all? The answer depends on whom you ask. *Meliorists* (Stanovich and West, 2000; Ariely, 2009; Kahneman, 2011) see the normative-descriptive gap as formidable, and a high level of human rationality as a rare phenomenon. From this viewpoint, being highly rational is like being a concert pianist—a great achievement and an unusual one. However, human rationality is amenable to education, and part of the Meliorist mission is to suggest how it might be improved. In contrast, *Panglossians* (Gigerenzer et al., 1999; Gladwell, 2007; Oaksford and Chater, 2009) see human rationality as a built-in evolutionary toolkit. Being rational, in this viewpoint, is the default—most of us are rational by dint of being human, just as most of us can all see and walk. If there is a gap between human behavior and any particular normative system, it is the normative system that is usually at fault.

As conflicting as they seem, Panglossianism and Meliorism nevertheless share some common ground. Both positions are *normativist:* They accept that rationality is measured by conformity to certain normative standards, while disagreeing, at least to an extent, on what those standards are, and how far the conformity exists. It is easy to see that identifying which normative standard is the right one would have far-reaching consequences for the Panglossians vs. Meliorists debate. Some normative standards may fit human behavior better than others, decreasing the normative-descriptive gap. In particular, the proponents of Bayesian rationality (Oaksford and Chater, 1998, 2007, 2009) suggested that probabilistic norms might provide a better fit to human rationality than norms derived from classical logic. However, arbitrating between normative standards is far from trivial. Elqayam and Evans (2011) criticized normativist theories (Panglossian and Meliorist alike) for trying to base this arbitration on empirical evidence, and so being in danger of committing the dubious inference from *is* to *ought,* considered a fallacy by many philosophers (Hudson, 1969; Pigden, 2010).

OPEN ACCESS

Edited and reviewed by:
Eddy J. Davelaar,
Birkbeck, University of London, UK

***Correspondence:**
Shira Elqayam
selqayam@dmu.ac.uk

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *27 March 2016*
Accepted: *14 April 2016*
Published: *28 April 2016*

Citation:
Elqayam S and Over DE (2016) Editorial: From Is to Ought: The Place of Normative Models in the Study of Human Thought. Front. Psychol. 7:628. doi: 10.3389/fpsyg.2016.00628

Normativist stances and practices are not, however, a universal phenomenon across the cognitive sciences. Linguists, for example, tend to be a lot less worried about violations of normative rules. Indeed, a tradition going back to De Saussure (1966) explicitly eschews normative concerns in favor of focusing on descriptive rules of language, the internalized ones that native speakers have in their heads. In the psychological literature on moral judgment, attitudes are more mixed, perhaps because the normative status of moral guidelines is far more controversial (although see Sunstein, 2005, for an attempt to derive moral norms from behavior).

The more recent position of *descriptivism* in the rationality debate (Elqayam and Evans, 2011; Evans and Elqayam, 2011; henceforth, collectively E&E) aims to follow in the footsteps of the Saussurean revolution in linguistics, proposing that the psychology of reasoning and decision making would be better off letting go of normative concerns altogether. Instead of measuring rationality by normative standards, the descriptivist position is that rationality should be measured by the achievement of personal goals. Evans and Over (1996) made a relevant distinction here between rationality$_1$, measured by achieving one's goals, and rationality$_2$, measured against some given normative standards. People can be rational$_1$ without being rational$_2$ and often are; and it is rationality$_1$ that is basic. Rationality$_1$ is personal, contextualized, and relative, resulting in *grounded rationality* (Elqayam, 2012).

This Research Topic in *Frontiers in Cognitive Science* follows in the wake of a *Behavioral and Brain Sciences* treatment on normativism and descriptivism (E&E; and see commentaries there). In the current issue our aim was to widen the debate, allowing more space for discussion as well as empirical contributions. The result is a range of 23 articles from some 54 authors, on a diversity of topics from moral judgment to theory of mind. We divided the book into six main sections.

THE STANDARD PICTURE: NORMATIVIST PERSPECTIVES

This section encompasses contributions from the *standard picture* (Stein, 1996), that is, the classical normativist perspective. We start with Baron's introduction of the standard picture in the field of judgment and decision making (JDM). Because JDM is an applied field, normative models are necessary in order to evaluate behavior, with a view to ultimately improving it. In addition to normative and descriptive models, we also need *prescriptive* models, which specify how such improvement can be achieved.

Oaksford argues that rationality$_1$ and rationality$_2$ are inseparable. By Davidson's *charity principle*, rationality depends on normatively evaluating other people's behavior. Moreover, as logic and probability are compatible, there is no need to arbitrate between these normative standards. And, given that probabilistic norms are universal, the relativist concerns proposed in Evans and Elqayam (2011) and Elqayam (2012) are unjustified.

Hahn responds to three critiques of normative Bayesianism (Elqayam and Evans, 2011; Jones and Love, 2011; Bowers and Davis, 2012), arguing that the critique of Bayesian models is too general to be valid or useful. Specific accounts of reasoning, decision making, and argumentation provide counterexamples to the claim that Bayesian modeling is too flexible and thus un-falsifiable. Normative considerations have explanatory power that cannot be matched by descriptive accounts or process-level analysis.

Crupi and Girotto argue that what might appear to be debates about standards in classical reasoning and decision making are in fact nothing of the sort. Instead, the controversies are about mapping the stimuli or the responses onto specific norms, or a failure to identify what the relevant norm should be.

We conclude this section with Quintelier and Zijlstra's take on the is-ought problem itself. They suggest that an is-to-ought argument might actually be normative. They argue that inferring is to "ought" from "is" is best treated as a type of defeasible inference, rather than deductive inference. Such arguments should not be judged for their validity or soundness, but by appealing to the appropriate standards or evaluating defeasible arguments.

IN DEFENSE OF SOFT NORMATIVISM

Emerging from the debate in E&E, soft normativism is the view that, within boundaries, normative models have an important role to play in the psychology of reasoning and decision making, alongside more descriptivist considerations. Soft normativism comes with a moderate degree of relativism, which both contributions to this section accept. Stupple and Ball suggest that, as long as researchers are cautious of normativist research biases and focus on processing models, normative benchmarks have a role to play: they enrich our understanding of processing models, particularly in the Meliorist context of improving reasoning and judgment. Achourioti et al. draw on Searle's distinction between constitutive and regulative norms, arguing that normative models in reasoning and decision making are important for specifying both. The challenge in reasoning is to select the normative models appropriate to one's goals.

EXPLORING NORMATIVE MODELS

The three contributions in this section focus on exploring and defending specific normative models. Markovits draws on Inhelder and Piaget (1958) to defend classical logic as the preferred normative model of rationality, arguing that the developmental evidence supports a notion of validity based on the existence of counterarguments. In contrast, the other two contributions support alternative normative models. Pothos and Busemeyer propose *quantum probability* as superior in explanatory power to classic (Bayesian) probability. Schwartenbeck et al. explore the *free energy principle*.

DESCRIPTIVIST PERSPECTIVES

The contributors in this section accept the descriptivist position as departure point for their analysis. Evans takes descriptivism even further by arguing that the very notion of irrationality is

problematic, depending as it does on an illusory presupposition that people are in conscious control of their minds and decisions. He concedes that deviations from normative standards can—and should—be construed as errors. However, these errors are merely evidence for limited capacity rather than irrationality.

Two further contributions explore normativism and descriptivism beyond the psychology of reasoning and judgment, ranging into Theory of Mind territory in philosophy and psychology. Iijima and Ota focus on the Knobe effect (Knobe, 2003), in which judgments of intentionality are affected by the perceived morality of the action. E&E argue that philosophers often misinterpret the Chomskyan distinction between competence and performance as normative, a muddle going back to Cohen (1981). Iijima and Ota accept this critique and extend it further, to criticize experimental philosophers for trying to draw unwarranted normative conclusions from the Knobe effect. Lastly, Wilkinson points out to the normativist stance in the psychological study of folk psychology. She argues that over-focus on questions of right and wrong in this study holds back research, and that more attention to the processing mechanisms underlying folk psychology would benefit the field.

EVOLUTIONARY AND ECOLOGICAL ACCOUNTS

Much of the rationality debate in reasoning and decision making is cast in evolutionary, adaptationist, and ecological terms (Over, 2003), with approaches ranging from massive modularity (Cosmides and Tooby, 1994), through fast and frugal heuristics (Gigerenzer et al., 1999), to dual processing and beyond (Stanovich, 2004). This section presents three such accounts. We start with Brase's massively modular account. Like Achourioti et al. Brase rejects the one-size-fits-all notion of rationality. Instead, in each modular domain, a different type of rationality predominates, linked to the evolutionary goals set by the domain—such as self-protection, mate acquisition, and kin care, among others.

If Brase considers evolutionary pressures a source of rationality, Goel sees them as the opposite. Like Evans, Goel highlights the role of implicit, unconscious sources of thinking and deciding, but unlike Evans, he regards them as prime examples of irrationality. He argues that neither massive modularity nor dual processing accounts provide adequate explanations for the universal biological cues that trigger irrational behavior. Instead, he proposes an *adulterated rationality* account, in which a late-evolving rational system is often inadequately equipped to suppress instinctual, irrational responses.

Lastly for this section, Schurz presents a two-dimensional charting of cognitive success. First, he points to a parallel between the normative/instrumental rationality distinction and the *deontological*/*consequentialist* distinction in meta-ethics. In each, the basis of justification is either *a priori* normative obligation, or the utilitarian consequences of one's actions, respectively. The distinction, between *a priori* intuitions and *a posteriori* success, akin to Gigerenzer's ecological rationality, is orthogonal to the one between *logically-general* accounts and *locally-adaptive* ones. This two-dimensional mapping of rationality gives rise to novel research questions, supporting a dual account of cognition in which the selection of the appropriate cognitive tool takes center stage.

EMPIRICAL REPORTS

In this last and largest section we present a collection of empirical reports, with methods ranging from simulation to modeling. The *new paradigm* in psychology of reasoning (Elqayam and Over, 2013), a Bayesian and decision theoretic approach to reasoning, is strongly in evidence here. The section launches with two new paradigm studies of conditional reasoning, both using everyday causal conditionals such as "If oil prices continue to rise, then UK petrol prices will rise." In both papers, descriptive models are held to provide a better fit for the data than models based purely on normative distinctions. Singmann et al. found that conformity, above chance, to coherence in conditional reasoning depends on the form of the inference. They advocate the *dual source* theory as a descriptive model. *This contribution was awarded the Best Student Paper Award of the Priority Program "New Frameworks of Rationality" for 2015.*

Trippas et al. used SDT (signal detection theory) to fit a large dataset of causal conditional reasoning with ROC (receiver operating characteristics) curves. They found that the descriptive theoretical modeling based on the difference between denial and affirmation inferences provided a better theoretical fit then the normative model based purely on inference validity. A debate follows this contribution, in which Singmann and Kellen dispute the SDT modeling in Trippas et al. arguing that they failed to make an unambiguous distinction between argument strength and response bias. Trippas et al. respond with a justification of their methods, and hold that their original point, that normative accounts are unreliable guides to conditional reasoning, remains in force.

Klaczynski uses individual differences measures to predict normative responding. Drawing on the classic methods of dual processing theories, he presents a large-scale individual differences study, showing that numeracy only predicted performance for participants who were both cognitively able and cognitively motivated.

We conclude this section—and the Topic—with two studies of judgment in social contexts. Wenmackers et al. simulation study identifies the individualistic nature of traditional approaches to human rationality, criticizing them for failing to take into account social-epistemic interactions between agents, such as information exchange. They test and support the Hegselmann–Krause model of epistemic interactions using computer simulations, which they argue are a useful bridge between normative models and descriptive results. Lastly, Gold et al. take the discussion (as Schurz does) into the realm of moral judgment. They criticize the artificiality of the trolley problems so widely used in moral judgment studies. Using more realistic scenarios, both in hypothetical contexts and operationalized in real life, they found that utilitarian responses were judged as more

morally right for actors than for onlookers in traditional trolley problems, but reverse was true for a hypothetical game show context. When the game show was enacted in real life, the results reverted to the trolley dilemma pattern. They conclude with a discussion of the design choices in moral judgment experiments.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Ariely, D. (2009). *Predictably Irrational, Revised and Expanded Edition: The Hidden Forces that Shape Our Decisions*. New York, NY: HarperCollins.

Bowers, J. S., and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* 138, 389–414. doi: 10.1037/a0026450

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford, UK: Oxford University Press.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behav. Brain Sci.* 4, 317–370. doi: 10.1017/S0140525X00009092

Cosmides, L., and Tooby, J. (1994). Beyond intuition and instinct blindness: toward an evolutionary rigorous cognitive science. *Cognition* 50, 41–77. doi: 10.1016/0010-0277(94)90020-5

De Saussure, F. (1966). *Course in General Linguistics* (original publication 1916). New York, NY: McGraw-Hill.

Elqayam, S. (2012). Grounded rationality: descriptivism in epistemic context. *Synthese* 189, 39–49. doi: 10.1007/s11229-012-0153-4

Elqayam, S., and Evans, J. St. B. T. (2011). Subtracting 'ought' from 'is': descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Elqayam, S., and Over, D. E. (2013). New paradigm psychology of reasoning: an introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Think. Reason.* 19, 249–265. doi: 10.1080/13546783.2013.841591

Elqayam, S., Thompson, V. A., Wilkinson, M. R., Evans, J. S. B. T., and Over, D. E. (2015). Deontic introduction: a theory of inference from is to ought. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 1516–1532. doi: 10.1037/a0038686

Evans, J. St. B. T., and Elqayam, S. (2011). Towards a descriptivist psychology of reasoning and decision making. *Behav. Brain Sci.* 34, 275–290. doi: 10.1017/S0140525X11001440

Evans, J. St. B. T., and Over, D. E. (1996). *Rationality and Reasoning*. Hove: Psychology Press.

Gigerenzer, G., Todd, P. M., and The ABC Research Group (1999). *Simple Heuristics that Make us Smart*. New York, NY; Oxford: Oxford University Press.

Gladwell, M. (2007). *Blink: The Power of Thinking without Thinking*. London: Penguin.

Hudson, W. D. (ed.). (1969). *The Is-Ought Question: A Collection of Papers on the Central Problem in Moral Philosophy*. London: Macmillan.

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking*. New York, NY: Basic Books.

Jones, M., and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? *On the explanatory status and theoretical contributions of Bayesian models of cognition. Behav. Brain Sci.* 34, 169–188. doi: 10.1017/S0140525X10003134

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis* 63, 190–194. doi: 10.1093/analys/63.3.190

Oaksford, M., and Chater, N. (1998). *Rationality in an Uncertain World*. Hove: Psychology Press.

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M., and Chater, N. (2009). Précis of bayesian rationality: the probabilistic approach to human reasoning. *Behav. Brain Sci.* 32, 69–120. doi: 10.1017/S0140525X09000284

Over, D. E. *(ed).* (2003). *Evolution and the Psychology of Thinking: The Debate*. Hove: Psychology Press.

Pigden, C. R. (2010). *Hume On is and Ought*. New York, NY: Palgrave Macmillan.

Searle, J. R. (2005). What is an institution? *J. Inst. Econ.* 1, 1–22. doi: 10.1017/S1744137405000020

Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago: Chicago University Press.

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate. *Behav. Brain Sci.* 23, 645–726. doi: 10.1017/S0140525X00003435

Stein, E. (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford: Oxford University Press.

Sunstein, C. R. (2005). Moral heuristics. *Behav. Brain Sci.* 28, 531–542. doi: 10.1017/S0140525X05000099

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The Standard picture: Normativist perspectives

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 24 December 2012
doi: 10.3389/fpsyg.2012.00577

The point of normative models in judgment and decision making

Jonathan Baron*

Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA
**Correspondence: baron@psych.upenn.edu*

Edited by:
Shira Elqayam, De Montfort University, UK

In this comment, I shall try to summarize arguments that I have made before (Baron, 1985, 1994, 2004, 2006, 2008). These arguments are my attempt to state the standard view in the field of judgment and decision making (JDM).

JDM is applied psychology. The ultimate goal is to improve judgments and decisions, or keep them from getting worse. In order to achieve this goal we need to know what good judgments and decisions are. That is, we need criteria for evaluation, so that we can gather data on the goodness of judgments, find out what makes them better or worse, and test method for improving them when there is room for improvement. This is the main function of normative models.

Examples of normative models in JDM are:

1. For quantitative judgments (e.g., populations of cities, proportions of coin tosses that were heads): the normative model is simply the right answers. This also applies to relative judgments (which city has more people?) or judgments of category membership. We can also quantify departures from the right answers in various ways.
2. For judgments of the probability of unique events, one type of normative model, which is applied to a group of such judgments, scores the judgments by distance from 0 (no) or 1 (yes) and applies some formula to these scores. A related approach is to aggregate judgments with the same stated probability (e.g., all those with 80%), and ask if the proportion is correct (calibration, the proposition should be true 80% of the time).
3. Alternatively, for probabilities of related unique events, we can assess their coherence, their agreement with each other. If you say that the probability is 0.6 that X will win a competition and 0.7 that Y will win, you are not coherent.
4. For decisions, we can sometimes assess their consistency with basic principles of decision making, such as dominance (if A is better than B in some respects and worse in no respects, then choose A).
5. More typically, we assess the coherence of sets of decisions, using a mathematical model to define coherence, such as expected-utility theory or exponential discounting (for decisions over time). "Utility" is a summary measure of "good(ness)."

We could, in principle, define normative models in terms of the behavioral steps involved in making a good judgment or decision. For example, we could define the normative model for subtraction problems in terms of the steps of subtracting digits, regrouping, etc. But, as just illustrated, most normative models in JDM do not do this and are thus not computational, in the sense of being specified as procedures.

Note that some normative models concern coherence of responses with each other while others concern correspondence with the world, a distinction made first by Hammond (1996) [see Dunwoody (2009), for an overview]. Correspondence-type models are usually difficult to apply to decisions, so that are used mostly for judgments. This because the "right answer" to a decision question usually depends on the values of the decision maker.

JDM makes distinctions among three types of models: normative, descriptive, and prescriptive. The three-way distinction emerged clearly in the 1980s (Freeling, 1984; Baron, 1985; Bell et al., 1988—all of whom wrote independently of each other), although various parts of it were implicit in the writing of Herbert Simon and many philosophers (such as J. S. Mill).

Normative models, as noted, are standards for evaluation. They must be justified independently of observations of people's judgments and decisions, once we have observed enough to define what we are talking about. When not obvious, as in the case of simple correspondence (the "right answer"), they are typically justified by philosophical and mathematical argument (Baron, 2004). Particularly in cases where we want to quantify deviations from the single best response, several normative models may apply to the same case (e.g., scoring rules for probability judgments).

Descriptive models are psychological theories that try to explain how people make judgments and decisions, typically in the language of cognitive psychology, which includes such concepts as heuristics and strategies, as well as formal mathematical models. Within the three-model framework, descriptive models are most useful when they explain departures from normative models, so researchers often focus on the search for such explanations. Such models allow us to determine whether, and, if so, how, we might improve judgments and decisions. When a deviation from a normative model is found to be systematic, not just the result of random error, we call it a bias. For example, people are biased to choose default options, even when others are normatively equal or better.

Prescriptive models are designs for improvement. If normative models fall in the domain of philosophy (broadly defined) and descriptive models in the

domain of empirical psychological science, then prescriptive models are in the domain of engineering (again, broadly defined). Originally, they were conceived as including mathematical tools that were useful for the formal analysis of decisions. These constitute the field of decision analysis, which includes several methods (and which has a society and a journal by that name). But prescriptive models can also be educational interventions (Larrick, 2004), which, for example, teach people alternative heuristics, to counteract heuristics that lead to biases.

A recent addition to the arsenal of prescriptive methods is the idea of "decision architecture" (Thaler and Sunstein, 2008), which consists of designing the presentation of decisions to those who will make them in such a way as to help people make the normatively better choice. A classic example is using the fact that people are biased toward the default to help them choose wisely by making what is usually the wise choice the default. For example, use a diversified portfolio as the default retirement plan for new employees (as opposed to, say, shares in company stock).

Thus, the ideal plan for JDM, sometimes actually realized (Baron, 2008; Thaler and Sunstein, 2008), is to apply normative models to judgments and decisions, looking for possible biases, then use the tools of psychology to understand the nature of those biases, and then, in the light of this understanding, develop approaches to improve matters. Of course, in real life these steps are not sequential, but are informed by each other. For example, decision analysis turns out to require the measurement of personal probability and utility, so now a large descriptive and normative enterprise is devoted to this measurement problem, which has produced better methods for measurement, which, in turn, are used to improve the original prescriptive models.

This plan clearly requires that the three elements are kept distinct. Suppose, for example, we make arguments for normative models on the basis of (descriptive) observations of what people do, under the assumption that people are rational. Then, we are likely to conclude that people are rational and that no prescriptive interventions are needed. The field of JDM would tend to disappear. Arguably, economics as a field made this assumption of rationality and thus was never concerned with helping people to make better economic choices, until recently, when economics has started to take the findings of JDM very seriously.

Another danger that JDM tries to avoid is to design prescriptive interventions without at least some clarity about normative and descriptive models. Specifically, we try to avoid "fixing things that ain't broke." This sort of prescription has happened in psychology. For example, it was assumed that creativity was limited by a lack of divergent thinking ("thinking outside the box"), and many programs to improve creativity assumed this, despite the fact that the evidence indicate quite clearly that this was not a common problem [e.g., Johnson et al. (1968); and see Perkins (1981), for an overview].

Much of the debate within JDM is about the seriousness of various purported biases. Although strong advocates on one side or the other tend to think either that people are hopelessly biased or that we are perfectly adapted to our environment, more moderate folks think that, while it all depends on the person, the situation, and the task, there really are some situations where people can be helped, sometimes a lot, through the JDM approach (Thaler and Sunstein, 2008).

We need to keep normative and prescriptive models separate as well. If we assume that normative models are also prescriptive, they may become self-defeating. In decision making, the main normative standard is the maximization of (expected) utility, and the time required for calculation usually reduces utility. If normative models require elaborate calculation, then, when a real person attempts to apply one to a decision, the utility loss from the time spent may be greater than the gain from using the model, as opposed to some simpler heuristic. In many cases, then, normative models are applied by researchers, and real people may use various heuristics to improve their judgments as evaluated by the normative models (e.g., Davis-Stober et al., 2010).

On the other hand, summary versions of normative models may require no calculation at all and may serve the purpose of focusing attention on only what is relevant. For example, utilitarianism, a variant of utility theory that applies to decisions that affect many people, says that the goal of such decisions is to maximize total utility. A real person can often *save* time by simply asking, "Which option produces the best outcome on the whole, considering effects on everyone?" (Baron, 1990). Such a question is often easy to answer, and it can avoid more elaborate reasoning when, for example, this simple principle is must be weighed against another, non-utilitarian, principle such as "Do not use one person as a means to help another." This conflict may occur in decisions about whether to abort a fetus, which would die anyway, in order to save the mother's life. When the fetal death is caused by abortion, then it is a means, and Catholic moral doctrine has been interpreted as prohibiting abortion for this reason, despite its obvious utilitarian benefit. The utilitarian solution is simpler because it involves only one principle and the decision maker does not need to resolve the conflict with another.

REFERENCES

Baron, J. (1985). *Rationality and Intelligence*. New York, NY: Cambridge University Press.

Baron, J. (1990). Thinking about consequences. *J. Moral Edu.* 19, 77–87.

Baron, J. (1994). Nonconsequentialist decisions (with commentary and reply). *Behav. Brain Sci.* 17, 1–42.

Baron, J. (2004). "Normative models of judgment and decision making," in *Blackwell Handbook of Judgment and Decision Making*, eds D. J. Koehler and N. Harvey (London: Blackwell), 19–36.

Baron, J. (2006). President's column: normative, descriptive, and prescriptive. Newsletter of the Society for Judgment and Decision Making, December. Available online at: http://www.sjdm.org/newsletters/06-dec.pdf

Baron, J. (2008). *Thinking and Deciding, 4th Edn*. New York, NY: Cambridge University Press.

Bell, D. E., Raiffa, H., and Tversky, A. (eds.). (1988). *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. New York, NY: Cambridge University Press.

Davis-Stober, C. P., Dana, J., and Budescu, D. V. (2010). Why recognition is rational: optimality results on single-variable decision rules. *Judgm. Decis. Mak.* 5, 216–229.

Dunwoody, P. T. (2009). Introduction to the special issue: coherence and correspondence in judgment

and decision making. *Judgm. Decis. Mak.* 4, 112–115.

Freeling, A. N. S. (1984). A philosophical basis for decision aiding. *Theory Decis.* 16, 179–206.

Hammond, K. R. (1996). *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavailable Injustice*. New York, NY: Oxford University Press.

Johnson, D. M., Parrott, G. L., and Stratton, R. P. (1968). Production and judgment of solutions to five problems. *J. Edu. Psychol. Monogr. Suppl.* 59, 1–21.

Larrick, R. P. (2004). "Debiasing," in *Blackwell Handbook of Judgment and Decision Making*, eds D. J. Koehler and N. Harvey (London: Blackwell), 316–337.

Perkins, D. N. (1981). *The Mind's Best Work*. Cambridge, MA: Harvard University Press.

Thaler, R. H., and Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Received: 04 December 2012; accepted: 07 December 2012; published online: 24 December 2012.

*Citation: Baron J (2012) The point of normative models in judgment and decision making. Front. Psychology **3**:577. doi: 10.3389/fpsyg.2012.00577*

This article was submitted to Frontiers in Cognitive Science, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

PERSPECTIVE ARTICLE
published: 15 May 2014
doi: 10.3389/fpsyg.2014.00332

Normativity, interpretation, and Bayesian models

Mike Oaksford*

Department of Psychological Sciences, Birkbeck College, University of London, London, UK

Edited by:
David E. Over, Durham University, UK

Reviewed by:
Ulrike Hahn, University of London, UK
Vincenzo Crupi, University of Turin, Italy

****Correspondence:***
Mike Oaksford, Department of Psychological Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK
e-mail: mike.oaksford@bbk.ac.uk

It has been suggested that evaluative normativity should be expunged from the psychology of reasoning. A broadly Davidsonian response to these arguments is presented. It is suggested that two distinctions, between different types of rationality, are more permeable than this argument requires and that the fundamental objection is to selecting theories that make the most rational sense of the data. It is argued that this is inevitable consequence of radical interpretation where understanding others requires assuming they share our own norms of reasoning. This requires evaluative normativity and it is shown that when asked to evaluate others' arguments participants conform to rational Bayesian norms. It is suggested that logic and probability are not in competition and that the variety of norms is more limited than the arguments against evaluative normativity suppose. Moreover, the universality of belief ascription suggests that many of our norms are universal and hence evaluative. It is concluded that the union of evaluative normativity and descriptive psychology implicit in Davidson and apparent in the psychology of reasoning is a good thing.

Keywords: psychology of reasoning, Bayesian models, Bayesian argumentation, radical interpretation, Donald Davidson, evaluative normativity

Elqayam and Evans (2011) have argued against *evaluative* normativity having any role in psychological theories of reasoning. They contrast evaluative normativity with *directive* normativity. They argue that directive normativity is conditional and perfectly consistent with programs in cognitive science like rational analysis (Anderson, 1990; Oaksford and Chater, 1998, 2007). Consequently, they have no problem with formulations like, *if you want to be well adapted to your environment then you should act in a Bayes optimum fashion in classification, decision and prediction.* However, what we can't apparently assert is the unconditional *you should act in a Bayes optimum fashion in classification, decision, and prediction.* This is an evaluative claim suggesting in some absolute sense that this is the right way to behave. In particular, they observe that if there were an alternative normative theory of what constitutes being well adapted to your environment, citing empirical evidence to distinguish between these two normative theories would commit the *is-ought* fallacy. Consequently, evaluative normativity should be expunged from psychological theorizing about reasoning.

In this paper, I pursue a broadly Davidsonian (Davidson, 2004) response to Elqayam and Evans' (2011). In the first section, *Types of Rationality*, I set up the argument by observing that two distinctions they make, between *instrumental* and *normative* rationality and between *directive* and *evaluative* rationality, are far more permeable than they require. I conclude that Elqayam and Evans (2011) primary objection is to the suggestion that we should pick the theory that makes the most rational sense of our data. In the second section, *Interpretation, Argumentation, and Rationality*, I argue that this is inevitable consequence of Davidson's account of radical interpretation. On Davidson's view, rationality is a social construct where to interpret others' statements requires that we adopt a principle of charity, i.e., they share the same norms as ourselves. Davidson's account suggests attributing people with intentional states like beliefs requires evaluative normativity. I then show that in the social context of argumentation, a third person argument evaluation methodology yields close conformity to rational Bayesian norms. Participants are quite capable of evaluating others arguments. I conclude that this ubiquitous human behavior is something that psychology must explain. In the final section, *How Many Rational Norms Are There*? I argue that logic and probability theory are not really competing norms, the important psychological question is whether beliefs are binary or graded. Moreover, following Davidson, I question Elqayam and Evans (2011) grounds for normative relativism. In conclusion, I suggest that while there are many outstanding problems and exceptions, the continuing union of evaluative normativity and descriptive psychology apparent in the psychology of reasoning is a good thing.

TYPES OF RATIONALITY

Stanovich (2011) argued that Elqayam and Evans (2011) drive a wedge between Bayesian probability theory, which they regard as an account of normative rationality, and *instrumental* rationality. Instrumental or practical rationality, which Elqayam and Evans (2011) endorse, provides a suitable means for achieving one's goals regardless of the nature of those goals. However, as Stanovich (2011) observes, this is a difficult wedge to drive home given that the standard justification for the laws of subjective probability are given by the Dutch book theorem (Vineberg, 2011). For each of the laws of probability theory, this theorem establishes that violating them would leave an agent open to making bets they cannot win. The *converse* Dutch book theorem then establishes that these laws are instrumentally rational because conforming to them prevents taking self-defeating actions. This instrumentally rational justification can then be provided with a directively rational formulation: *if an agent wishes to avoid making bets they*

cannot win, then they should conform to the laws of probability theory. This conditional formulation just restates the *converse* Dutch book theorem. So this formulation involves making conformity to the normative theory conditional on that normative theory's rational justification. The justification for probability theory is instrumental [other epistemic justifications, based on maximizing accuracy, are equally instrumental (Joyce, 1998)]. So, in the case of probability theory there is simply no wedge to be driven between instrumental and normative rationality[1].

This formulation also raises the question of how universal are the goals stated in the antecedent? In a conditional formulation the more universal an antecedent the less it needs to be stated. So, for example we would normally say *ripe apples fall*. We do not feel compelled to formulate this as *if gravity is in force ripe apples fall*. One could even use an appropriate modal, *ripe apples ought to fall*. Certainly one might be inclined to query whether this is a real or a good apple if it did not fall, which is perilously close to an evaluative judgment. Similarly, the more universal we regard the wish to avoid making bets one is bound to lose, the more inclined we would be to drop the conditional formulation and evaluate anyone not conforming to the rules of probability as irrational just as we may be inclined to evaluate the apple as inedible. If we encountered someone willing to make bets they were bound to lose, they would probably be institutionalized for their own safety. As with instrumental and normative rationality, the barrier between directive and evaluative rationality seems permeable. Moreover, the fundamental issue is of universality versus relativity. The theory is normatively rational if its justification is considered universal.

The inference to which Elqayam and Evans (2011) seem to take exception is the claim that as theoreticians we *should* accept the theory that makes the most rational sense of the participants' behavior (Oaksford and Chater, 1996, 2007). As long as we are comparing the rules of normative theories, this will mean that the one that best describes participants' behavior is the one that makes most rational sense of it. This thesis derives from the fact that in interpreting empirical data, i.e., our participants' behavior, we are in exactly the same position as the *radical* interpreter in Davidson's (1984, 2004) theory of ascribing intentional content. The difference is that as reasoning researchers we may have more than one normative theory in mind, whereas in radical interpretation one imputes one's own norms to one's interlocutor. However, the general principle remains the same: we are trying to make the best sense of what we have been told.

INTERPRETATION, ARGUMENTATION, AND RATIONALITY

Davidson's model of radical interpretation is an idealized account of how a cognitive agent can interpret another agent's behavior and utterances to infer their beliefs and desires (Rescorla, 2013). The model is based on Bayesian decision theory, in which beliefs are graded and related to subjective probabilities and people's desires are represented as utilities. Savage's (1954) axioms show that when a person's preferences meet certain requirements there are probabilities and utilities that guarantee that their preferences maximize expected utility. Consequently, an agent's beliefs and desires can be inferred from their overt preferences. An important wrinkle is that the propositional content of beliefs are not pre-specified but are also inferred from an interlocutor's preferences for the truth of sentences. Central to this account is the thesis that to ascribe another person with the appropriate beliefs and desires means we must assume they conform to our own standards of rationality. This is the principle of charity. As Davidson (2005; p. 319, cited in, Rescorla, 2013) puts it: "Charity is a matter of finding enough rationality in those we would understand to make sense of what they say and do, for unless we succeed in this, we cannot identify the contents of their words and thoughts." Rationality is constitutive of having intentional states.

This is an idealized model but the central idea that we must attribute to others similar rational norms to ourselves in order to interpret them is intended as a more general claim about interpretation in the real world that involves attributing others with propositional attitudes like beliefs and desires. On Davidson's view describing somebody's behavior in terms of beliefs and desires is inseparable from normative evaluation.

Davidson's (2004) emphasis on interpretive communicative processes proposes a particular research methodology which has been pursued recently in the context of human argumentation (Oaksford and Hahn, 2004, 2013; Hahn and Oaksford, 2007). Argumentation is a social phenomenon in which one or more people attempt to persuade another person or group of a particular, often controversial, position. It is a commonplace of argumentation theory that arguing is pointless unless there is broad agreement between the protagonists on what could count as a reasonable argument (Perelman and Olbrechts-Tyteca, 1969; Woods et al., 2004). Without this point of departure there is no point in engaging in an argumentative exchange. At least initially, we must apply the principle of charity[2]. Recently it has been argued that reasoning usually has an argumentative goal (Hahn and Oaksford, 2007; Mercier and Sperber, 2011). Consequently, it is in social argumentative contexts where people's rational norms of reasoning would be expected to be most in evidence. It is a critical ability to be able to evaluate the arguments put forward by others to persuade you or your friends of particular positions.

Recent research in this area has adopted a third person argument evaluation methodology (Oaksford and Hahn, 2004, 2013; Hahn and Oaksford, 2007; Harris et al., 2012). Participants are explicitly asked to assess the degree to which one interlocutor, A, *should* be convinced by an argument put forward by another interlocutor, B. So, participants are explicitly asked for an evaluative judgment. They are also provided with information about A's prior degree of belief in the conclusion. Hahn and Oaksford (2006, 2007), Oaksford and Hahn (2004, 2013) have provided normative Bayesian analyses of a variety of different forms of argumentation

[1]We note also that the justification for selecting data in accordance with Oaksford and Chater's (1994) information gain model is again instrumental. So following its dictates will mean that this strategy minimizes the length of the sequential sample needed for the posteriors to converge on the true hypothesis (Fedorov, 1972). This is an instrumental justification: *if* people want to get to the truth in the most economical way they will select data in accordance with the theory.

[2]After an initial exchange, we may discover that we are not in a critical discussion, i.e., a rational exchange of arguments intended to persuade, but rather are in a quarrel, where rationality goes out the window.

which make clear predictions for participants' judgments. In this context, a normative Bayesian account provides excellent fits to the data. Moreover, this is true even when there are no parameters free to vary (Harris et al., 2012) because participants have been asked for their judgments of the relevant likelihoods from which predictions for their posteriors can be directly computed (see also, Fernbach and Erb, 2013). These results demonstrate that when participants are asked for an evaluative judgment of other peoples' arguments they reveal behavior that is closely in accordance with the appropriate normative theory. This is not only because they have been asked directly to make an evaluative judgment. They are also explicitly provided with A's prior degree of belief, which absolves them from the dilemma of considering whether *they* would believe the conclusion prior to hearing the argument. They are simply told that, for whatever reason, A believes it to a certain degree. In first person paradigms, participants are asked to assume or suppose that *they* believe the premises to be true or to a certain degree, when of course they may believe no such thing.

In summary, the psychology of reasoning will have to deal with evaluative normativity because much human behavior involves the explicit evaluation of others' arguments, especially in politics, and in the law. Moreover, participants in experiments on argumentation make these evaluations naturally and their performance reveals direct sensitivity to appropriate rational norms.

HOW MANY RATIONAL NORMS ARE THERE?

I conclude this paper by addressing two critical issues underlying Elqayam and Evans (2011) criticisms of evaluative normativity, (i) deciding between normative theories and (ii) the conviction that constructs like the principle of charity collapse into relativism. On Davidson's (2004) ideal model there are no alternative normative frameworks. Basic logic, probability theory, and decision theory [see, Chater and Oaksford (2012) on the role of these theories in cognitive science] are fundamental rational norms and he broaches no other possibilities. This raises the question, of how many rational norms are there actually to choose between? A prima facie argument can be made that that there are not as many as one might think. Elqayam and Evans (2011) suggest that the new Bayesian paradigm is an alternative norm account. I argue that since probability theory presupposes standard logic they are not really in competition. A derived theorem of the Kolmogorov axioms is *logical consequence*, i.e., *if X logically entails Y, then* $Pr(Y) \geq Pr(X)$, which "ensures that probabilistic reasoning respects deductive logic" (Joyce, 2004, p. 135). The question is not whether one norm supplants another but whether beliefs are graded. Once we opt for graded beliefs, then we need to know how they are updated in inference when new information comes in. This can be achieved by Bayesian conditionalization rather than modus ponens (Oaksford, in press; Oaksford and Chater, 2007, 2013), although this is not necessary because probabilistic premises will *deductively* entail a probability interval for the conclusions of an argument (Pfeifer and Kleiter, 2010). Consequently, I suggest that the move to Bayesian probability is not a move to an alternative norm rather than a move to a finer grained analysis of beliefs which is not just binary true or false.

Thus, when comparing logic and probability, we are not choosing between competing norms. Davidson would argue, and common sense seems to dictate, that if the more nuanced view provides a rational understanding of more of the data it is the preferred theory. When the issue of competing norms is taken out of the equation this is simply the question of which theory provides the best description of the data. What happens if there are genuinely competing normative theories that are equally descriptively adequate?

For example, in decision theory an explicit competitor to classical Bayesian probability theory has been provided by quantum probability (Pothos and Busemeyer, 2013). This would appear to be much closer to the competing norms case that Elqayam and Evans (2011) envisage. Quantum probability stands to quantum logic – in which the law of the excluded middle is not valid – as Bayesian probability stands to standard logic (Oaksford, 2013). Moreover, across a variety of tasks, Pothos and Busemeyer (2013) argue that quantum probability is more descriptively adequate than Bayesian probability theory. Recall that the formulation for directive normativity is conditional, with the relevant justification for the normative theory in the antecedent. For Bayesian probability theory we have, *if an agent wishes to avoid making bets they cannot win, then they should conform to the laws of probability theory*. For quantum probability, however, there does not appear to be a relevant justificatory antecedent. There would appear to be no Dutch book theorem showing that failure to conform to the laws of quantum probability would lead anyone to make bets they could not win[3]. Moreover, conformity to the laws of quantum probability may well lead to a Dutch book being made against you. For example, it has been shown that committing the conjunction fallacy (Tversky and Kahneman, 1983) can allow a Dutch book to be made against you (Gilio and Over, 2012; Hahn, 2014) and quantum probability apparently *predicts* the conjunction fallacy (Pothos and Busemeyer, 2013). Consequently, however descriptively adequate with respect to the data quantum probability appears to be, it cannot explain how behavior succeeds in the real macroscopic world which we inhabit. Even if we can make sense of laying bets on the outcomes of quantum events, there would still need to be an independent argument that there are similar events about which we could gamble at the macroscopic level (Oaksford, 2013).

Elqayam and Evans (2011) argue against the principle of charity solely on the observation that norms are relative to particular cultural and historical contexts. However, they do not discuss Davidson's view of rationality as a constitutive norm (Rescorla, 2013). On Davidson's view conformity to these norms is constitutive of having intentional states and is not relative to any particular cultural or historical context. As there are no human beings to whom we would not attribute beliefs this suggests that our norms are also universal. The Dutch book theorems certainly have this character. Gambling is a universal human activity, engaged in by

[3]Although in physics, there are arguments that a Bayesian approach, i.e., probability as a measure of ignorance, might make sense of quantum probability as a theory of rational betting in quantum gambles (Pitowsky, 2003). One then has to ask whether there is any analog of a quantum gamble at the macroscopic level that any human being would be concerned to win.

all cultures and in all historical contexts. Moreover, it seems inconceivable that anyone would fail to accede to the rationale for the Dutch book theorems, what normal human being would wish to make bets they are bound to lose? In the first section, I argued that the permeability between directive and evaluative rationality depends on the universality of the justification for a normative system. So we have good grounds to view probability theory as a universal evaluative norm.

CONCLUSION

In conclusion, in the psychology of reasoning, interpreting experimental results, just as in interpreting another's utterances, requires making the best rational sense of the observed behavior. People evaluate each other's arguments in politics and in the law and in appropriate argumentative contexts their judgments conform to the rational norms of probability theory. The current Bayesian turn in the psychology of reasoning addresses the question of whether beliefs are graded and is not an alternative norm to standard logic. From Davidson's perspective, the universal attribution of beliefs to others has the corollary that our rational norms are likely to be similarly universal. Elqayam and Evans (2011) provide no grounds to question this perspective. However, there are many exceptions, data that does not conform to these norms (e.g., Tversky and Kahneman, 1983; but see, Crupi et al., 2008), cases of irrationality due to illness or injury, cases where sacred values are opposed to utility maximization (Atran and Axelrod, 2008), and other paradoxes of maximizing expected utility (Burns and Wieth, 2004; but see Turner and Quilter, 2014). However, there are responses to these exceptions as some of the citations indicate. In sum, the union of evaluative normativity and descriptive psychology, implicit in Davidson (Rescorla, 2013), is continuing to yield important results and this should be regarded as a good thing.

REFERENCES

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Mahwah, NJ: Erlbaum.

Atran, S., and Axelrod, R. (2008). Reframing sacred values. *Negotiation J.* 24, 221–246. doi: 10.1111/j.1571-9979.2008.00182.x

Burns, B. D., and Wieth, M. (2004). The collider principle in causal reasoning: why the monty hall dilemma is so hard. *J. Exp. Psychol. Gen.* 133, 434–449. doi: 10.1037/0096-3445.133.3.434

Chater, N., and Oaksford, M. (2012). "Normative systems: logic, probability, and rational choice," in *The Oxford Handbook of Thinking and Reasoning,* eds K. Holyoak and R. Morrison (Oxford: Oxford University Press), 11–21.

Crupi, V., Fitelson, B., and Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Think. Reason.* 14, 182–199. doi: 10.1080/13546780701643406

Davidson, D. (1984). "On the very idea of a conceptual scheme," in *Inquiries into Truth and Interpretation,* ed. D. Davidson (Oxford: Oxford University Press), 183–198.

Davidson, D. (2004). *Problems of Rationality.* Oxford: Clarendon Press. doi: 10.1093/0198237545.001.0001

Davidson, D. (2005). *Truth, Language, and History.* Oxford: Clarendon Press. doi: 10.1093/019823757X.001.0001

Elqayam, S., and Evans, J. T. (2011). Subtracting 'ought' from 'is': descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Fedorov, V. (1972). *Theory of Optimal Experiments.* London: Academic Press.

Fernbach, P. M., and Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1327–1343. doi: 10.1037/a0031851

Gilio, A., and Over, D. (2012). The psychology of inferring conditionals from disjunctions: a probabilistic study. *J. Math. Psychol.* 56, 118–131. doi: 10.1016/j.jmp.2012.02.006

Hahn, U. (2014). The Bayesian boom: good thing or bad? *Front. Psychol.* (in press).

Hahn, U., and Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese* 152, 207–236. doi: 10.1007/s11229-005-5233-2

Hahn, U., and Oaksford, M. (2007). The rationality of informal argumentation: a Bayesian approach to reasoning fallacies. *Psychol. Rev.* 114, 704–732. doi: 10.1037/0033-295X.114.3.704

Harris, A. L., Hsu, A. S., and Madsen, J. K. (2012). Because hitler did it! Quantitative tests of Bayesian argumentation using ad hominem. *Think. Reason.* 18, 311–343. doi: 10.1080/13546783.2012.670753

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philos. Sci.* 65, 575–603. doi: 10.1086/392661

Joyce, J. M. (2004). "Bayesianism," in *The Oxford Handbook of Rationality,* eds A. R. Miele and P. Rawling (Oxford: Oxford University Press), 132–155. doi: 10.1093/0195145399.003.0008

Mercier, H., and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–74. doi: 10.1017/S0140525X10000968

Oaksford, M. (2013). Quantum probability, intuition, and human rationality. *Behav. Brain Sci.* 36, 303. doi: 10.1017/S0140525X12003081

Oaksford, M. (in press). "Knowing enough to achieve your goals: Bayesian models and practical and theoretical rationality in conscious and unconscious inference," in *Human Rationality: Thinking Thanks to Constraints,* eds L. Macchi, L. M. Bagassi, and R. Viale (Cambridge, MA: MIT Press).

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 608–631. doi: 10.1037/0033-295X.101.4.608

Oaksford, M., and Chater, N. (1996). Rational explanation of the selection task. *Psychol. Rev.* 103, 381–391. doi: 10.1037/0033-295X.103.2.381

Oaksford, M., and Chater, N. (eds). (1998). *Rational Models of Cognition.* Oxford: Oxford University Press.

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning.* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198524496.001.0001

Oaksford, M., and Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Think. Reason.* 19, 346–379. doi: 10.1080/13546783.2013.808163

Oaksford, M., and Hahn, U. (2004). A bayesian analysis of the argument from ignorance. *Can. J. Exp. Psychol.* 58, 75–85. doi: 10.1037/h0085798

Oaksford, M., and Hahn, U. (2013). "Why are we convinced by the ad hominem argument?: Bayesian source reliability and pragma-dialectical discussion rules," in *Bayesian Argumentation,* ed. F. Zenker (Dordrecht: Springer), 39–59.

Perelman, C., and Olbrechts-Tyteca, L. (1969). *The New Rhetoric: A Treatise on Argumentation.* Notre Dame, IN: University of Notre Dame Press.

Pfeifer, N., and Kleiter, G. (2010). "Mental probability logic," in *Cognition and Conditionals: Probability and Logic in Human Thinking,* eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 153–173. doi: 10.1093/acprof:oso/9780199233298.003.0009

Pitowsky, I. (2003). Betting on the outcomes of measurements: a Bayesian theory of quantum probability. *Stud. Hist. Philos. Mod. Phys.* 34, 395–414. doi: 10.1016/S1355-2198(03)00035-2

Pothos, E. M., and Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behav. Brain Sci.* 36, 255–274. doi: 10.1017/S0140525X12001525

Rescorla, M. (2013). "Rationality as a constitutive ideal," in *A Companion to Davidson,* eds E. Lepore and K. Ludwig (Oxford: Wiley-Blackwell), 472–488. doi: 10.1002/9781118328408.ch27

Savage, L. (1954). *The Foundations of Statistics,* 2nd Edn. New York: Dover.

Stanovich, K. E. (2011). Normative models in psychology are here to stay. *Behav. Brain Sci.* 34, 268–269. doi: 10.1017/S0140525X11000161

Turner, R., and Quilter, T. (2014). *The Two Envelopes Problem* (*Gatsby Centre for Computational Neuroscience*). Available at: http://www.gatsby.ucl.ac.uk/~turner/Notes/TwoEnvelopes/2envlps.pdf (accessed January 3, 2014).

Tversky, A., and Kahneman, D. (1983). "Extensional vs. intuitive reasoning: the conjunction fallacy in probability judgement," in *Heuristics and Biases: The Psychology of Intuitive Judgement,* eds T. Gilovich,

D. Griffin, and D. Kahneman (New York: Cambridge University Press), 19–48.

Vineberg, S. (2011). "Dutch book arguments," in *The Stanford Encyclopedia of Philosophy,* Summer 2011 Edn, ed. E. N. Zalta. Available at: http://plato.stanford.edu/archives/sum2011/entries/dutch-book/ (accessed January 4, 2014).

Woods, J., Irvine, A., and Walton, D. N. (2004). *Argument: Critical thinking, Logic and the Fallacies* (Revised Edition). Toronto, ON: Prentice Hall.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The Associate Editor declares that while the author Mike Oaksford as well as the reviewer Ulrike Hahn are currently affiliated with the same department (Department of Psychological Sciences, Birkbeck College, University of London), there has been no conflict of interest during the review and handling of this manuscript.

Received: 05 February 2014; accepted: 31 March 2014; published online: 15 May 2014.
*Citation: Oaksford M (2014) Normativity, interpretation, and Bayesian models. Front. Psychol. **5**:332. doi: 10.3389/fpsyg.2014.00332*
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 08 August 2014
doi: 10.3389/fpsyg.2014.00765

The Bayesian boom: good thing or bad?

Ulrike Hahn*

Department of Psychological Sciences, Centre for Cognition, Computation, and Modelling, Birkbeck, University of London, London, UK

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Mike Oaksford, University of London, UK
David E. Over, Durham University, UK
Jonathan St. B. T. Evans, University of Plymouth, UK

****Correspondence:***
Ulrike Hahn, Department of Psychological Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK
e-mail: u.hahn@bbk.ac.uk

A series of high-profile critiques of Bayesian models of cognition have recently sparked controversy. These critiques question the contribution of rational, normative considerations in the study of cognition. The present article takes central claims from these critiques and evaluates them in light of specific models. Closer consideration of actual examples of Bayesian treatments of different cognitive phenomena allows one to defuse these critiques showing that they cannot be sustained across the diversity of applications of the Bayesian framework for cognitive modeling. More generally, there is nothing in the Bayesian framework that would inherently give rise to the deficits that these critiques perceive, suggesting they have been framed at the wrong level of generality. At the same time, the examples are used to demonstrate the different ways in which consideration of rationality uniquely benefits both theory and practice in the study of cognition.

Keywords: Bayesian modeling, rationality, normativity, probability

INTRODUCTION

The last two decades of cognitive science have seen a bit of a revolution: probabilistic models of cognition, in particular, Bayesian models have not only steadily increased in volume, but have come to grab a large market share in those outlets, such as Psychological Review, that focus on psychological "theory." These trends are manifest not just in a wealth of reviews (e.g., Chater et al., 2006, 2010) and bibliometric statistics, but, last but not least, in the fact that Bayesian models have recently prompted a number of high-profile critiques (e.g., Elqayam and Evans, 2011; Jones and Love, 2011; Bowers and Davis, 2012a,b). A pre-requisite to critique is getting noticed in the first place, and, given that these critiques concern formal, mathematical models of cognition, that is no mean feat.

So these critiques may plausibly be taken to signal a moment of arrival in the development of the paradigm, particularly given that they were written for a general audience, not just for specialists within the discipline. At the same time, it seems likely that these critiques provide insight that research would be well-advised to heed. In light of this, the present paper scrutinizes these recent critiques with a view to identifying the key implications they present for future work.

FUNDAMENTAL CRITIQUES

Three sets of criticisms have recently been aimed at Bayesian models of cognition: the target article in Behavioral and Brain Sciences by Jones and Love (2011) raising the specter of "Bayesian fundamentalism," Bowers and Davis article in Psychological Bulletin (2012) on "Bayesian just-so stories" and, from an even broader perspective, Elqayam and Evans (2011) recommendation to abandon a central role for normative models in the study of the cognition. While there is some overlap between these critiques, each makes distinct points. Each is also a lengthy article in its own right, containing a wealth of observations and claims. However, for the purposes of this article, four main claims of interest will be highlighted and addressed for each.

JONES AND LOVE (2011)

Jones and Love find that rational Bayesian models are (1) significantly unconstrained, because they are generally uninformed by either process-level data or environmental measurement. Furthermore, (2) the psychological implications of most Bayesian models are also unclear (last but not least because there is little contact with mechanism or process). The retreat to the level of abstraction away from process at which Bayesian models are typically phrased is not perceived to be of intrinsic interest because (3) Bayesian inference itself is conceptually trivial (Bayes' theorem is just a simple "vote counting"). And finally, (4) many Bayesian models simply recapitulate existing (mechanistic level) theories.

BOWERS AND DAVIS (2012A,B)

Here it is maintained that (1) flexibility with priors, likelihoods, and utility functions frequently makes models unfalsifiable, while (2) Bayesian theories are also rarely better at predicting data than alternative (and simpler) non-Bayesian ones. In general, for understanding cognition and building insightful models of cognitive processes, (3) constraints other than rational analysis are more important. As a consequence, (4) psychology and neuroscience now abound with Bayesian "just so" stories, that is, mathematical analyses of cognition that can be used to explain almost any behavior as optimal.

ELQAYAM AND EVANS (2011)

The focus of Elqayam and Evan's critique, finally, is more general in its target than just Bayesian modeling, affecting also the use of decision-theory and logic as other putative norms of rationality. The central point in Elqayam and Evan's paper is (1) a critique of what they call "normativism": the idea that human thinking

reflects a normative system against which it should be measured and judged. Normativism is conceptually dubious because it invites fallacious is-to-ought and ought-to-is inferences (2). At the same time, little can be gained from normativism that cannot be achieved by descriptivist computational-level analysis (3). As a consequence, Elqayam and Evans believe that (4) theories of higher mental processing would be better off if freed from normative considerations.

Each of these articles has already seen extensive counter-critique, last but not least the open peer commentaries that are an integral part of the journal format for two of these three articles (and for the third, Bowers and Davis, 2012a,b), see the reply in the same journal by Griffiths et al., 2012). It is the contention of the present paper, however, that there are still things to be said on this topic, and that some things that have been said deserve to be said again and become clearer or more compelling when put together in a single overall argument. First and foremost, it is the contention of this paper that closer consideration of actual examples of Bayesian treatments of different cognitive phenomena allows one to defuse the above critiques. Specifically, it will be argued that one of the main reasons the critiques go amiss is that they have been phrased at the wrong level of generality. More detailed consideration of specific examples, however, is not something the restrictive format of open peer commentary readily supports.

THE DIVERSITY OF BAYESIAN MODELING

One of the tensions in all three critiques is that, while it is likely they have been motivated by particular applications, they are pitched as general critiques of a paradigm. This is striking because Bayesian probability itself is, in first instance, a formalism, that is, a "language." As such, it affords many and diverse applications. How then could such a diverse set of applications suffer from common problems? For one, it could do so coincidentally: researchers who avail themselves of this language happen to, by and large, be researchers who are comparatively poor at the task of model-building. For example, they may fail to appreciate fundamental criteria of "goodness" for a model that a field has managed to identify. The root cause, in this case, is effectively sociological. There is nothing within the formalism itself that makes necessary the deficits observed, and, in the hands of others, these limitations could easily be rectified. The second possibility is that there is some deeper limiting factor in the formalism that is responsible for the perceived limitations. In this latter case, the formalism itself is indeed, at least partly, to blame. Both cases would merit critique, but the nature of that critique, in order to be appropriate and hence constructive, would have to be very different. The only way to distinguish between these two possibilities is to consider specific examples. Limitations of the formalism itself should emerge as common aspects of all examples considered.

For these purposes it is important to consider a broad range of examples. **Figure 1** contains a set of such examples, chosen with diversity in mind. The list contains both some of the most famous and influential Bayesian modeling (e.g., Anderson, 1991; Oaksford and Chater, 1994) and other examples, which, by comparison, are completely obscure (e.g., Harris and Hahn, 2009). The examples vary also in the cognitive domain to which the model is applied, ranging from judgment through reasoning and argumentation to categorization and language acquisition.

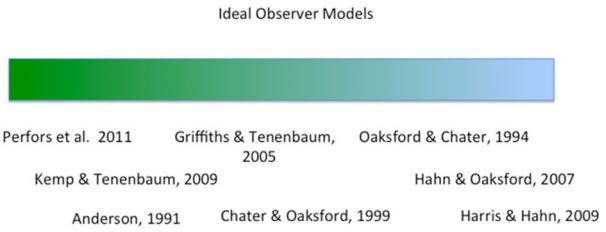


FIGURE 1 | The figure presents examples of applications of Bayesian modeling. The examples, discussed in the text, reflect both different types of application in terms of the aspect of the cognitive system modeled, and the theoretical and methodological role accorded to Bayesian inference as a result.

In fact, these differences in domain give rise to an informal ordering within the Figure: the green-blue dimension[1]. This dimension may, in first instance, be taken to reflect the extent to which the underlying cognitive task *inherently involves inference, and more specifically, probabilistic inference.*

To illustrate: On the far right hand end of the "blue spectrum," the task participants face in Harris and Hahn's (2009) studies of evidential coherence is that of evaluating, from the perspective of the police, the potential location of a body given the testimony of (less than perfectly reliable) multiple witnesses. Not only is this inherently an inferential task involving uncertainty, but participants are specifically asked to evaluate a question about 'how likely it is' that the body lies within a particular area on a map.

By contrast, at the green end, Anderson's (1991) famous rational model of unsupervised categorization addresses the task of imposing categories on unlabeled instances, that is, partitioning a set of objects into distinct classes of object. This need not be viewed as an inference task at all. Furthermore, even if the task is to be construed as one involving inference, there is a wealth of different choices concerning what that inference may be about. Ultimate answers to the fundamental question of what unsupervised categorization does and what it is for rest on extremely difficult questions about the relationship between mind and world (e.g., the extent to which we "discover" categories in the world or instead impose them) and the role of categories in language and thought.

In fact, rival accounts of unsupervised categorization which assume that classification proceeds on the basis of inter-item similarity, for example, may assume that such similarities reflect deep facts about the environment (or, human perceptions thereof, given that "similarity" is a subjective, not an objective relation between objects, see e.g., Hahn and Chater, 1997), or they may simply take as their point of departure that human categorization seems sensitive to similarity.

[1]Both are pleasing colors in keeping with the fact that the dimension does not reflect value.

Anderson's (1991) model is based on the idea that categorization reflects the goal of optimally predicting the unseen features of objects, that is, we wish to be able to predict $P_i(j|F_n)$, the probability that (as yet unseen) dimension i of the object possesses the value j, given the feature structure F_n observed so far. Categories are formed to assist this goal. Hence, objects are assigned to categories in such a way as to make the feature structures of those objects most probable. As a Bayesian model, the rational model assigns a new object to the most probable category k given the features observed, $P(k|F)$. In so doing, the model may choose to create an entirely new category for that item.

The fact that the two examples, Harris and Hahn's study of coherence, and Anderson's rational model, fall on opposite ends of the spectrum with regard to the extent to which the task under investigation is *necessarily* construed as involving probabilistic inference has immediate implications for the role of rational, Bayesian inference in each case.

Where the task is uncontroversially construed as an inferential one, the mapping between task and formalism is more or less direct. Where it is not, the probabilistic construal is merely one of many possible, equally plausible, task decompositions. This has direct consequences for the "normative" or "rational" status bestowed by Bayesian inference. While it is the case that Bayesian probabilistic inference has a privileged status that makes its use "rational" or "optimal" in certain well-defined senses (more on this in a moment), this normativity or rationality only goes as far as the inference itself. If the task may plausibly be construed as not involving inference in the first place, then the resultant model as a whole is neither inherently more "normative" or "rational" than any other.

Associated with the difference in role for Bayesian inference at the two ends of the green-blue spectrum are other differences. In Harris and Hahn's (2009) study prior probabilities are objectively defined within the task. There is nothing to "choose" here by the modeler, and there are no free parameters. In the case of Anderson's rational model, by contrast, model behavior is critically dependent on prior probabilities for category membership. Anderson (1991) specifies this prior in the following way:

$$p(k) = \frac{cn_k}{(1-c)+cn} \tag{1}$$

where n_k is the number of objects assigned to category k thus far, n is the total number of classified objects and c is the so-called "coupling parameter." This parameter governs the probability that a new instance will receive an entirely new label, $P(0)$:

$$p(0) = \frac{1-c}{(1-c)+cn} \tag{2}$$

In other words, the coupling parameter determines how readily new categories will be formed: for high values of the coupling parameter, larger clusters are favored by the prior, whereas for low values the model will favor greater numbers of smaller categories. Model behavior thus varies dramatically as a function of c.

Furthermore, the combinatorial explosion concerning the number of possible partitions of even fairly small sets of to-be-classified objects means that Anderson's model must rely on approximation to the optimal Bayesian estimates. Alternative approximation algorithms to Anderson's are possible (e.g., Gibb's sampling, see Geman and Geman, 1984) or particle filters (see e.g., Doucet et al., 2001), and, as Sanborn et al. (2010) demonstrate, will give rise to differences in model predictions.

This makes it fuzzier what the rational model actually *is*, and makes the model harder to test empirically. However, contrary to concerns about Bayesian models articulated by Bowers and Davis (2012a,b) there is no sense in which the rational model is unfalsifiable. One can readily evaluate model predictions across values of the coupling parameter and contrast those predictions with human behavior (as in Sanborn et al., 2010) and in that way compare the rational model with competing formal models of unsupervised categorization (as in Pothos et al., 2011), and one can do this for different approximation algorithms.

Needless to say, in the case that other models perform better on such tests (as Pothos et al., indeed find them to do), no one would take that to indicate that participants' views on classification are "irrational." Because there are so many ways the goals of categorization can be construed, the model does not prescribe what people *should* do in any strong sense. Deviating from it is not an "error" in the same way that prominent inferential failures in the judgment and decision-making literature (such as the conjunction fallacy, Tversky and Kahneman, 1983) are viewed as errors—an issue we return to below.

Concerning the critical challenge surrounding model falsifiability it seems important to distinguish vague predictions from model flexibility. Vagueness means that it is unclear exactly what predictions are, and what empirical evidence might or might not meet them. Flexibility, by contrast, means that a model or theory can change its predictions depending on parameterization; given a particular set of parameters, however, predictions are specific. The rational model not only has an important free parameter, but due to the nature of its approximation algorithm, also has stochastic variation in its model output; however, by averaging over model runs, specific predictions can be derived, and—as has been demonstrated empirically (see e.g., Pothos et al., 2011)—the model can readily be compared both with human data and with other models.

Beyond pointing out that even a flexible model such as Anderson's rational model admits falsification it is hard to know how to address Bowers and Davis claims that Bayesian models may frequently be unfalsifiable given their flexibility with priors, liklihoods and utility functions. It seems hard to see that Bayesian models are more flexible than other mathematical models that admit of parameterization. They are certainly not inherently more flexible, because in many contexts (certainly toward the "blue end" of **Figure 1**), Bayesian models of the task can and have been applied (and compared with human performance) without free parameters at all, because parameters such as priors or likelihoods are derived from participants estimates or because they are taken directly from environmental quantities and the model itself consists simply of Bayes theorem. In addition to the Harris and Hahn (2009) paper, other examples here include Harris et al.'s (2012) study on argumentation, and the extensive body of research within the 1960's that examined experimentally human belief revision using simple devices such as colored pokerchips

drawn from bags of varying chip composition (see e.g., Peterson and Beach, 1967; Edwards, 1968). At the very least, these examples make clear that the formalism itself does not impose any particular degree of flexibility.

Other examples along the green-blue dimension fit also with the first two examples of Anderson (1991) on the one hand, and Harris and Hahn (2009) on the other. Perfors et al. (2011) simulations are aimed at addressing fundamental questions in language acquisition concerning so-called poverty of stimulus arguments, that is, arguments that seek to argue that certain aspects of language, though developmentally acquired, cannot be learned, because there is insufficient information in the linguistic input to the child (for a review and references see also e.g., Hahn and Oaksford, 2008). Perfors et al. like many researchers concerned with these questions before them (see e.g., Chomsky, 1957, 1986; Gold, 1967; Wharton, 1974) assume that the task at hand is to infer a grammar, from which the grammatical sentences of the language can be generated. However, whether this is an appropriate way to conceive of language acquisition is in itself a matter of debate. Other researchers have argued that the goal of acquisition is to learn form-meaning mappings (e.g., Bates and MacWhinney, 1989) or to learn procedures for comprehension and production (Seidenberg and MacDonald, 1999). On such views, there need be no role at all in language for a grammar as traditionally conceived. The role of Bayesian inference in Perfors et al.'s study is thus to provide an elegant, well-defined, and well-understood modeling tool. The point is not an account of what children *should* do.

Over at the "blue end" of **Figure 1**, however, such normative concerns are integral to Oaksford and Chater's (1994) account of Wason's selection task, a paper that, like Anderson's rational model, is a cornerstone of Bayesian modeling. Wason's classic (1968) study shows participants deviating from a falsificationist strategy when asked to select information to test a rule. While falsification was advocated as an ideal strategy for science by Popper (1959), it is not ideal in general, that is, independently of the specific hypotheses and nature of the environment as shown, for example, by Klayman and Ha (1989). And indeed, philosophers of science have not only noted that falsificationism does not capture the actual conduct of science (Kuhn, 1962; Lakatos, 1976), but have moved away from it as an ideal strategy in more recent work that adopts a Bayesian, normative perspective on scientific inference (e.g., Earman, 1992; Howson and Urbach, 1993). Oaksford and Chater (1994) seek to show that under certain simple assumptions about the structure of the environment, and certain assumptions about reasonable priors, participants' responses on the selection task are well-understood as an approximation to optimal data selection.

In general, Oaksford and Chater's treatment of conditional reasoning involves a twofold argument. On the one hand, they argue that the utility of classical logic in the context of everyday reasoning is extremely limited (see e.g., Oaksford and Chater, 1991); probability theory, by contrast, provides a natural formalism for reasoning under uncertainty. On the other hand, as they seek to demonstrate, seeming patterns of deviation in human responding on what have traditionally been conceived of as logical reasoning tasks, are well-captured under the assumption that participants view the seemingly deductive inference task as a probabilistic inference task.

This work is naturally situated toward the "blue end" as it is concerned with what are inference tasks by design. There is room for debate here on a normative level about the mapping between probability theory and the task; in particular there has been considerable philosophical debate about the appropriate formalization of the natural language condition "if ... then" (see e.g., Edgington, 1995; Evans and Over, 2004), so the normative claims do not simply have to be accepted at face value. But they are integral to the overall aims of the project. At the same time, there is a descriptive component: the claim that actual participant responding is well-understood as an approximation of this normative construal. This descriptive claim may be empirically challenged, both by seeking to provide evidence of systematic deviation between model and observed behavior, and by positing alternative explanations of behavior that rest on functionally different interpretations (by participants) of the task.

Lively empirical debate has thus ensued (see e.g., the open peer commentary on Oaksford and Chater, 2009). This in itself testifies against claims about lack of falsifiability, but it is also important to note here that Oaksford and Chater's work has, in fact, brought a new level of specificity to behavioral prediction in the context of logical reasoning (see also Hahn, 2009 for discussion of this point). Prior to Oaksford and Chater's work, data in the psychology of logical reasoning were a collection of qualitative phenoma ("context effects," "supression effects" etc.). Since their seminal (1994) paper, empirical work in the psychology of reasoning frequently involves evaluation of detailed quantitative predictions. This was first seen in Oaksford and Chater's probabilistic approach, and it is "rival approaches" that have followed in this (see e.g., Schroyens and Schaeken, 2003; Oberauer, 2006; Klauer et al., 2007).

This example speaks to a whole range of separate points in the above critiques of Bayesian models: namely, the shift to more detailed, quantitative predictions provides a ready example where Bayesian models do not simply recapitulate existing mechanism level theories [Jones and Love (4) above]; moreover, it provides an example where a Bayesian model has been "better at predicting data than simpler (non-Bayesian) alternatives" [see, Bowers and Davis, (2) above]; and it makes questionable the claim that "normativism" has hampered the development of high-level cognition so that we would be better off without it [Elqayam and Evans, (3 and 4)], and that constraints other than rational analysis are more important [Bowers and Davis (3)].

It is precisely the fact that the Bayesian framework enables quantitative prediction that enabled Oaksford and Chater's work to bring about this change in specificity of prediction within the psychology of reasoning, and their choice of formalism was driven by normative considerations. Other quantitative models may have followed subsequently, but the impulse for the shift came from the use of Bayesian modeling.

It is worth emphasis also that the reasoning tasks addressed in Oaksford and Chater's work are classic examples of "high-level cognition" which Fodor (1983) considered to be "central processing," and hence an aspect of cognition for which we would never have detailed theories and predictions. That the field of reasoning can capture subtle changes in behavior in response to changes in the content of high-level, verbal experimental materials in such detail is thus, in and of itself, a remarkable success.

Moreover, Oaksford and Chater's treatment of selection task and logical reasoning (see also on syllogistic reasoning, Chater and Oaksford, 1999) are not alone here. Arguably, this specificity has been spreading through other aspects of human reasoning as well (see also e.g., Kemp and Tenenbaum, 2009). Hahn and Oaksford's work on informal argument fallacies are a further case in point (e.g., Hahn and Oaksford, 2007). Fallacies, or arguments that seem correct but aren't, pervade everyday informal argument. Catalogs of argumentation fallacies (also known as reasoning fallacies) originate with Aristotle and have been of concern to philosophers, logicians, and argumentation theorists to this day, though they have engendered only small amounts of psychological research in the past (e.g., Neuman and Weitzman, 2003). The longstanding goal of fallacies research has been to provide a comprehensive, formal treatment that can explain exactly why they are "bad" arguments. Hahn and Oaksford (2007) show how classic fallacies, such as the argument from ignorance ("ghosts exist, because nobody has proven that they don't"), or circular arguments ("God exists, because the Bible says so and the Bible is the word of God") can be given a formal Bayesian treatment that distinguishes appropriately weak examples of these argument forms from ones that seem intuitively acceptable. More generally, it provides explanations of widespread intuition that arguments from ignorance or circular arguments are frequently weak: analysis across the range of possible underlying probabilities that these arguments may involve demonstrates how they are typically weaker than other types of arguments in everyday life (for details see Hahn and Oaksford, 2007).

This is in part an explicitly normative project, aimed at addressing long standing theoretical questions about the fallacies, but also more general questions about the extent to which there can be "norms" for argument quality that allow us to determine whether an argument *should* or should not convince.

At the same time, the ability to measure argument quality through use of the Bayesian, probabilistic framework allows one to generate both qualitative and quantitative predictions against which people's judgments of everyday arguments can be compared. Such comparisons have been conducted, not just in the context of the fallacies, but in the context of other arguments as well (e.g., Hahn and Oaksford, 2007; Hahn et al., 2009; Corner et al., 2011; Harris et al., 2012).

The predictions made in these contexts are not only novel, there is, in many of the cases examined, simply no alternative framework that would allow one to make predictions about the materials examined[2]. That is, the theoretical questions that can be addressed are new. But there are not just new questions about how people evaluate particular argument forms which have now been formalized. The formal framework provides a methodological tool that allows one to examine a whole host of issues concerning argumentation that are not possible without it. For example, as Corner and Hahn (2009) note, much of the communication to the public of socio-scientific issues of broad concern such as climate change, genetically modified foods, nanotechnology and so on, involves brief summaries of arguments. How people evaluate such arguments is thus a central practical concern across a broad range of issues requiring large-scale action. A normative standard for measuring argument quality, and with that participants' evaluation of arguments, provides a tool for probing whether the way people think about issues such as climate change (for example with respect to conflicting testimony, see e.g., Lewandowsky et al., 2013) differs from the way they reason in other evidential contexts. Such comparisons become possible despite the differences in argument content (and hence attendant differences in people's prior beliefs and the actual diagnosticity of the evidence) because responses to arguments from different domains can be compared *via the normative standard*: in other words, one can ask whether people's reasoning is more or less in line with normative prescriptions across different domains.

Far from re-capitulating the predictions of other, simpler, or more process-oriented models, then, this argumentation work has created a wealth of opportunity for empirical inquiry. Against the claim that other computational level theories might be as successful (or even more successful) if the limiting emphasis on normative considerations were abandoned stands the simple fact that no other computational level theory presently exists in this particular case. Given the fact that the development of the computational level theory was driven explicitly by normative considerations, it would also seem perverse to consider such considerations a block to progress [cf. Elqayam and Evans (4)], at least in this context.

Similarly, the argumentation example is at odds with the perception that "other kinds of constraints" (e.g., neural constraints) are, typically, more powerful than rational or normative considerations. And this seems indicative of "the blue end" of **Figure 1** more generally. For example, it is a characteristic of Oaksford and Chater's work in the psychology of reasoning that it is precisely not concerned with process or implementation. Greater predictive power with regard to human behavior (i.e., the initial shift from qualitative to quantitative prediction) was achieved in their work despite moving to a higher level of abstraction. Moreover, the argumentation example may lead one to suspect that it is not despite that retreat to a higher level of abstraction but rather precisely because of it, that detailed quantitative predictions suddenly become possible.

What Bayesian modeling captures in this context is *relationships between information states*. If human reasoning and inference about the world is to have any point at all, it must be sensitive to the actual content of what is under consideration. Where evidential and inferential relationships are at stake, information content is the first and primary consideration. It is thus no coincidence that a probabilistic framework (which is about content) does a better job of predicting human behavior than the limited structural considerations of classical logic, for example. Of course, it is clear that reasoning will also be influenced by the mechanisms through which it is carried out. However, were these mechanisms to provide greater constraints on, say argument evaluation, than the actual information content of the argument and the relationship of that content to other beliefs, then these mechanisms would necessarily be extremely restricted inferential devices. Our best

[2]This is, of course, not to say that there has been no empirical work on other aspects of the fallacies or on argumentation more generally (for a recent overview see, Hahn and Oaksford, 2012).

evidence concerning higher level-cognition suggests that this is not what human thought is like[3].

For sure, there are deviations from "normative responding" in any reasoning or evidence evaluation context that has been examined, but the deviations would have to outweigh the correspondence to provide greater, more fundamental, and more useful initial constraints. Otherwise, starting from considerations of normative responding will provide the single biggest gain in predictive accuracy. Moreover, via inspection of systematic deviations, it likely provides one of the most powerful routes to identifying where mechanism constraints must be playing a role, and thus to what those mechanisms might be.

WHY NORMATIVE, WHY RATIONAL?

For many applications of the Bayesian framework the appeal to its normative status is integral. What then does that status rest on, and what kind of rationality or optimality can it consequently bestow?

In fact, there are multiple, independent routes to establishing a normative basis for Bayesian inference (see e.g., Corner and Hahn, 2013 for detailed discussion both of the general issue of normativity and Bayesian inference specifically). Lack of awareness of these distinct possibilities makes it easy to underestimate both the ways in which Bayesian inference may be perceived to provide a norm, that is a prescription of how one *ought* to behave, and to over-estimate how readily alternatives may make a rival claim. At the same time, lack of care in considering exactly what the normative status pertains to runs the risk of overblown normative claims for Bayesian models.

Of the different routes for claiming a normative basis for the Bayesian framework, the Dutch Book argument is the most well-known. A Dutch Book is a combination of bets that can be shown to entail a sure loss. In other words, engaging in a combination of bets that constitute a Dutch Book means necessarily incurring a loss, regardless of how the world turns out. Moreover, this loss is immediate, arising the moment the bet is resolved, not just in the long run (as incorrectly stated in Pothos and Busemeyer, 2013).

The Dutch Book argument provides an instrumental argument for assigning degrees of beliefs in accordance with the probability calculus based on the minimal assumption that incurring a sure loss would be undesirable. Specifically, the argument connects degrees of belief to a (theoretical) willingness to bet by assuming that a person with degree of belief P in a proposition a would be willing to pay up to £P to bet on a. The Dutch Book Theorem states that if a set of betting prices violates the probability calculus, then there is a Dutch Book consisting of bets at these prices, that is, a combination of bets that guarantees a sure loss. Being in possession of degrees of belief that violate the probability calculus makes possible Dutch Books and conversely, conformity with the calculus provides immunity from Dutch Books (the so-called converse Dutch book theorem, see e.g., Hajek, 2008).

Bayesian inference (and Bayesian modeling), however, is not just characterized by assignment of probabilities in line with the axioms of probability theory, but also by the use of Bayesian conditionalization for belief revision. That is, Bayes' theorem (which itself follows from the axioms of the probability calculus) is used as an update rule to accommodate new evidence. Analogous, so-called diachronic Dutch book arguments exist for Bayesian conditionalization (see Teller, 1973; and for the converse Dutch book argument, Skyrms, 1993).

To illustrate the nature of Dutch Book arguments with a famous example: Assigning to the conjunction of two events or claims a higher probability (or degree of belief) than is assigned to the less probable of the two—the so-called conjunction fallacy—is, in effect, a logical error. The conjunction of two events, A and B, cannot be true without each of the events being true also, and the event "A and B" cannot occur without the event A and the event B occurring as well. Hence they cannot be *less* probable than the conjunction; failing to realize this makes one Dutch-bookable, as exemplified in **Table 1** (see also Newell et al., 2007 for a concrete numerical example). For example, believing it to be more probable that Linda is a bankteller and a feminist, than that she is a feminist (Tversky and Kahneman, 1983) means that a combination of bets could be offered which, if accepted, would imply a sure loss.

The example of the conjunction fallacy is chosen here, in part, because it has been argued recently within the cognitive literature that quantum probability may provide a more appropriate framework for modeling human cognition than classical probability (e.g., Pothos and Busemeyer, 2009; Busemeyer et al., 2011). This not only involves the use of quantum probability as a descriptive tool, but its proponents have specifically asked about its normative or rational status (see e.g., Busemeyer and Bruza, 2012; Pothos and Busemeyer, 2013, 2014). For the conjunction fallacy, the ability to model what, from the perspective of classical logic and probability, are viewed as "errors" has been presented as one of the key modeling "successes" within the quantum framework (but see for challenges to its descriptive adequacy e.g., Tentori and Crupi, 2013). However, adherence to quantum probability in this way licenses the conjunction fallacy, and hence, is Dutch-bookable[4]. The Dutch book illustrates why this has traditionally been viewed as a mistake.

Unsurprisingly, in seeking to make their case for "quantum rationality," Busemeyer and colleague are skeptical about Dutch book arguments and the extent to which they justify a normative status for classical probability. In particular, they highlight a supposed practical limitation of Dutch Book justification: "Avoiding a Dutch book requires expected value maximization, rather than expected utility maximization, that is, the decision maker is constrained to use objective values rather than personal utilities, when choosing between bets. However, decision theorists generally reject the assumption of objective value maximization and instead allow for subjective utility functions (Savage, 1954).

[3]Even for a very restricted inferential device, however, there must be constraints on how its outputs "cohere" with those of other components of the systems if the system is to function effectively. This need for coherence once again brings a focus on information content and with it, a role for Bayesian inference (see Griffiths et al., 2012).

[4]In this application of quantum probability to a macro-level entity such as Linda the feminist bankteller. Needless to say, this is not the standard domain of application for the formalism.

Table 1 | Dutch book arguments.

The typical way to present Dutch Books is by presenting propositions, associated betting odds, and outcomes in a table. The left most example in the table below illustrates a bet on *a* for an agent who buys a bet with stake 1$ (i.e., 1$ is the amount won if *a* is true) for the price *q*(a) (*q* as in betting "quotient"); by assumption, the agent's betting quotient is determined by her degree of belief that *a* is true. The table is read in the following way: in the case where *a* turns out to be true, the agent receives 1$ as a payout, but has paid *q*(*a*) for the bet, so her net payoff is *1$ - q(a)*. If *a* turns out to be false, there is no payout, and the agent has simply lost the money she paid for the bet. She will make a profit if *a* turns out to be true and she has paid less than 1$ for the bet (i.e., *q(a)* < 1), and a loss otherwise.

a	**Net payoff**
True	*$1 - q(a)*
False	*−q(a)*

a	***b***	**Net payoff**	
True	True	*1 −q(a, b)*	***q(b)−1***
True	False	***−q(a,b)***	*q(b)*
False	True	***−q(a,b)***	***q(b)−1***
False	False	***−q(a,b)***	*q(b)*

The right hand of the table shows a Dutch Book for the conjunction fallacy. Here, *a* and *b* represent two claims, with *b* representing the less probable of the two. Our agent will *sell* for price *q(b)* a bet that pays out 1$ if *b* turns out to be true, and pay out 0 if it is false. Our agent will also *buy* for price *q(a,b)* a bet that pays out 1$ if the conjunction *(a,b)* is true and 0 otherwise. Because our agent commits the conjunction fallacy *q(a,b)* is greater than *q(b)*. In each row, the net payoff is negative, so whatever the truth or falsity of *a* and *b*, our agent makes a loss. This can be read off directly for rows 2–4 (quantities in bold are "losses," quantities in plain font are "gains"). In the case of row 1, where both *a* and *b* are true, our agent wins 1$ because the conjunction is true. From this 1$, the price paid for the bet needs to be deducted to calculate net gain. Against this is then set the loss the agent makes by paying out on the win for *b*. This loss necessarily exceeds the gains. (For two positive numbers *x* and *y*, if $x > y$, then $1-x < 1-y$; also, $y-1 = -(1-y)$; because *q(ab)* > *q(b)* by definition, the gain $1-q(ab)$ must be smaller than the loss $q(b)-1$, meaning a net loss overall).

This is essential, for example, in order to take into account the observed risk aversion in human decisions (Kahneman and Tversky, 1979). When maximizing subjective expected utility, CP [insertion: CP = Classical Probability] reasoning can fall prey to Dutch book problems (Wakker, 2010)" (Pothos and Busemeyer, 2013, p. 270).

This argument (largely repeated in Pothos and Busemeyer, 2014) conflates two separate issues: whether or not utilities are "subjective" and whether or not an agent is "risk averse." On the issue of subjective utilities and Dutch books, Pothos and Busemeyer are wrong: The Dutch Book argument could equally be run over subjective utilities (see e.g., Hajek, 2008). In general, the so-called representation theorems for expected utility[5] are typically defined over preferences- that is subjective valuations (see e.g., Karni, 2014). These representation theorems establish that as long as an agent's preferences respect certain fundamental axioms an expected utility representation of those preferences (which casts them as a combination of probability and utility) is guaranteed. Hence economists long assumed that people's choices might be well-described as "maximizing subjective expected utility." In their *descriptive* application of expected utility theory, they have also sought to allow for the fact that people are frequently "risk averse": many might, for example prefer 10$ for sure, over a 50/50 chance of receiving either 30$ or 0$, even though the expected value of the latter option is higher (namely 15$) and picking it will lead to greater gains on average.

Within Expected Utility Theory (EUT) risk aversion can be modeled by assuming that people have non-linear, concave utility functions whereby twice as much money becomes less than twice as "good"[6]. This does not mean that people *should* have non-linear utility functions and be risk averse, however. From the perspective of EUT, risk aversion *costs money*, and the degree to which the concave utility function diverges from a risk neutral, linear, utility function captures an agent's "risk premium," that is, the price an agent is willing to pay in exchange for certainty over and above expected monetary value. Given that risk aversion implies loss relative to expected value the possibility of Dutch Books under risk aversion seems unremarkable and simply highlights, in a different way, the cost of risk aversion. Risk aversion as a descriptive fact about human preferences does not make a Dutch Book a "good thing"; rather there may be practical contexts in which the price of susceptibility to Dutch Books may be a price an agent is willing to pay in exchange for some greater good. It is thus unclear how risk aversion undermines the Dutch Book argument.

Pothos and Busemeyer's argument is in many ways illustrative of the lively debate about Dutch book arguments. Such debate has focussed to a good extent on how literally one may interpret them and thus how far exactly is their normative reach (for extensive reviews see e.g., Hajek, 2008; for summaries of the main lines of argument see e.g., Corner and Hahn, 2013): for example, one can also avoid a particular Dutch book simply by refusing to bet (though we cannot refuse to bet against nature in general, i.e., we are forced in daily life to make decisions under conditions of uncertainty).

Such arguments do not detract from the fact that the existence of a Dutch book highlights a defect of sorts in a set of probabilities or degrees of belief (e.g., the failure to recognize that if the conjunction is true, each of the conjuncts is necessarily true also). And the defect highlighted (via the theoretical "sure loss")

[5]These are themselves often used as justifications for a normative basis of probability, see e.g., Armendt (1993).

[6]Though whether this is descriptively adequate seems doubtful, see e.g., Rabin and Thaler (2001).

is one that obtains regardless of the way the world is, that is, what actually turns out to be true or false.

Normative justification for Bayesian probability can thus also be derived from considerations of accuracy (examples of this are Rosenkrantz, 1992; Joyce, 1998; Leitgeb and Pettigrew, 2010a,b). Accuracy-based justifications involve the use of a scoring rule to measure the accuracy of probabilistic forecasts as used, for example, in meteorology, (e.g., Winkler and Murphy, 1968). Scoring rules allow one to assign credit for correct predictions, and penalties for incorrect ones. Overall accuracy is then reflected in the total score. Rosenkrantz (1992) shows that updating by Bayes' rule maximizes the expected score after sampling; in other words, other updating rules will be less efficient in the sense that they will require larger samples, on average, to be as accurate. This holds for any way of measuring accuracy that involves a so-called "proper scoring rule," that is, a scoring rule which will yield highest scores when agents report "honestly" their actual degrees of belief (that is, there is no incentive for agents to, for example, "hedge their bets" by reporting more conservative estimates than they believe). Furthermore, this optimality of Bayesian conditionalization with respect to maximizing accuracy holds not just for "interest-free inquiry," but also holds where actions dependent on our beliefs about the world are at stake: using Bayesian conditionalization to update our beliefs upon having sampled evidence maximizes expected utility (Brown, 1976; Rosenkrantz, 1992). Finally, Leitgeb and Pettigrew (2010b) demonstrate that for a common measure of accuracy (the Brier score, Brier, 1950), Bayesianism (i.e., assignment of probabilities in accordance with the probability axioms and updating via Bayes' rule) follows from the simple premise that an agent ought to approximate the truth, and hence seek to minimize inaccuracy. Being Bayesian will minimize inaccuracy of the agent's beliefs across all "possible worlds" the agent is conceptually able to distinguish and hence, in principle, to entertain![7]

These results provide a normative justification that, unlike the Dutch book argument, is direct: it is the goal of Bayesian inference to make inductive inferences about the world, and such inference is optimal in a well-defined sense, whereby—on average—no other procedure can do better.

What is true of induction in general, of course, can also be applied to specific cases. For example, in the context of supervised categorization, that is, the task of trying to assign instances, including novel instances, to the right (pre-existing) category, the so-called Bayes' optimal classifier will assign items to categories in such a way as to minimize the expected error rate, and thus provides a point of comparison in machine learning contexts (see e.g., Ripley, 1996)[8].

Considering in such detail various strands of justification for why "being Bayesian" might be viewed as normative or rational is important for a number of reasons. Vis a vis a "normative challenge" such as that by proponents of quantum probability, it makes clear quite how much is required for such a challenge to be well-supported. Merely assuming or speculating that human behavior is rational will never suffice to make it so, and Elqayam and Evans (2011), in particular, have been right to highlight that such an inference from "is" (i.e., how people behave) to "ought" (i.e., how they should behave) would be fallacious [see Elqayam and Evans (2) above]. However, the normative status of Bayesian probability does not rest on its descriptive fit to human behavior, but rather on independent arguments such as those just described.

Furthermore, it is because of these normative foundations, that Bayes' theorem, though conceptually simple, is far from conceptually trivial in the way Jones and Love (2011) might be taken to suggest (3 above). It figures centrally within formal work in the philosophy of science and within epistemology that is concerned with fundamental questions about information seeking, evidence, and explanation, and it figures centrally in statistics, machine learning and artificial intelligence (and that fact, incidentally, adds an interdisciplinary richness to Bayesian models both at the "blue" and the "green" end). For all these disciplines, normative questions about how one ought to behave, or how a problem is best solved, are both theoretically interesting and practically important. Indeed, the debate about Bayesian models itself is a debate about what should count as a "good" theory and about how psychological research "ought" to proceed.

It is thus an interesting question in and of itself how a particular model or procedure relates to an optimal Bayesian one. As a consequence, the theoretical interest and explanatory power of a Bayesian formalization does not rest on whether or not it makes deviant (and hence unique) predictions from existing psychological theories. Contrary to Jones and Love's critique that Bayesian models frequently merely recapitulate extant (mechanism level) theories (2 above) and to Bowers and Davis perception that they rarely "make better predictions" of human behavior than simpler, non-Bayesian models, there may be added value in "mere recapitulation" because it is informative with regard to normative concerns, which in turn opens up the possibility of functional explanations with regard to *why* the system is operating the way it does.

Of course, as outlined earlier in the context of Anderson's rational model, the normative force of Bayesian conditionalization applies only to the extent that Bayesian inference has a clear mapping onto the task under which it is a core component. Where it does, however, viewing a Bayesian formalization and a mechanistic model simply as "competitors" partly misses the point. Furthermore, the normative aspect may give Bayesian formalization a unique role in deriving adequate mechanistic accounts in the first place, as the final section of this paper will seek to show.

THE FALSE TENSION BETWEEN MECHANISM, PROCESS MODELS AND NORMATIVE ACCOUNTS

Running through the critiques of Bayesian modeling that form the focus of the present paper seems to be a perception that "rational" or "normative" considerations are blind to, or even at odds, with mechanism and process-level concerns; however, it may be argued that they are, in fact, part of the route to identifying mechanism or process-level constraints in the first place.

[7] With the proviso that these possible worlds are finite, a restriction that seems fine for creatures with finite resources and life spans.

[8] Consideration of the optimal Bayes classifier also makes clear that the "rational" force of Anderson's (1991) model increases the more one is willing to view the task of unsupervised categorization as one of discovering underlying, true categories in nature.

Specifically, it seems likely that pinning down properly cognitive constraints will require appeal to optimality. As Howes et al. (2009) have recently argued, the space of possible cognitive theories is massively under-constrained. The notion of *cognitively bounded rational analysis* provides a means by which to limit that search space in ways that other approaches do not allow, thus providing an essential complement to other methods. Specifically, the study of cognition faces the particular difficulty of humans' inherent flexibility: multiple strategies are typically available for any given task, and the project of seeking to discern cognitive invariants must distinguish between aspects of behavior that appear universal because they, in fact, reflect hard constraints within the system, and those that arise time and again simply because they reflect selection of an obvious, best strategy.

In light of this difficulty, Howes et al. (2009) demonstrate how making strategies computationally explicit, determining their expected pay-offs, and seeking to understand performance relative to those optimal strategies is fundamental to tackling the credit-assignment problem between "fundamental cognitive constraint" and "strategy selection."

Such an approach seems at odds with critiques of Jones and Love (2011), Elqayam and Evans (2011), and Bowers and Davis (2012a,b). In arguing that process level theories are more important and should be given precedence or that research would advance more quickly without normative theories, these critiques are overlooking the methodological value that stems from the fact that optimal models (in general) form a privileged class of explanation. It is a reasonable default assumption that the cognitive system is trying to do something sensible. Consequently, the fact that a strategy would be optimal supports a presumptive inference to the fact that it is indeed the strategy being used and this has been seen as methodologically important not just in psychology, but also economics and the social sciences.

The standard method of economics has been founded on optimization: Individual agents are presumed to be rational and it is the goal of economic theorizing to understand aggregate behaviors that arise from the interactions of such individuals (see e.g., Lehtinen and Kuorikoski, 2007). Rational choice theory has assumed that economic agents have stable and coherent preferences as set out by expected utility theory (Von Neumann and Morgenstern, 1947). This methodological commitment, though challenged by behavioral economics (see e.g., Thaler and Mullainathan, 2008), has not only been seen as successful within economics, but has been exported to adjacent disciplines such as political science (see e.g., Cox, 1999; Ferejohn, 2002).

Though conceived primarily as a normative theory, expected utility theory has, at times, been viewed as a descriptive theory within economics (see e.g., Friedman and Savage, 1948), and its normative appeal has been viewed as a prima facie reason for why it might provide a descriptive account (Friedman and Savage, 1952 see also Starmer, 2005 for critical discussion). Even now, given overwhelming evidence of violations of rational choice theory in both experiments and field studies (see e.g., Camerer, 1995), the theories of aggregate behavior arising from idealized rational agents aim to be descriptively accurate; this may be possible because certain behavioral contexts provide pressures that lead individuals to utility maximizing behavior (see e.g., Binmore, 1994; Satz and Ferejohn, 1994) and because the behavior of aggregate systems may be robust to the deviations from rational choice theory real agents might display (Lehtinen and Kuorikoski, 2007)[9]. None of this involves a fallacious *ought-to-is* or *is-to-ought* inference of the kind Elqayam and Evans accuse "normativism" of [see Elqayam and Evans (2) above]. Such a fallacy would be commited if one thought the world was a particular way simply because it ought to be, or, conversely, that something out to be the case simply because it was. However, the expectation of rational behavior simply thinks it *likely* that people behave a certain way because they ought to, not that they necessarily do; at the same time, what counts as rational does not rest on whether or not people actually behave the way they should (*is-to-ought*), because the normative claim has been independently derived[10].

More generally, rational standards provide essential interpretative tools: Any human behavior typically allows many different interpretations, and this is as relevant to science as it is to everyday life. In day-to-day life we resolve ambiguity with "the principle of charity" (e.g., Govier, 1987; see also Oaksford, 2014). Specifically, given multiple interpretations of what someone is saying, we pick the interpretation that renders what they are saying most sensible as our default interpretation. This interpretation may be wrong, and further evidence will force us to abandon it. However, the basic fact that there are default orderings over possible interpretations simplifies massively the task of understanding. Even without specific knowledge of an individual we can typically make reasonably accurate predictions just on the basis of what would be "sensible" (though again, there is no guarantee that these predictions will be correct).

The principle of charity likewise applies to the formal context of understanding behavior within psychological research (see also Hahn, 2011). If we observe something counter-intuitive or surprising, we should as researchers always ask ourselves whether there is an interpretation of participants' behavior that might render it sensible (and hence predictable). Such consideration may identify discrepancies in the way experimenter and participant view the task, leading the researcher to revise interpretations of what it is participants are doing, and many of the seeming "errors" and "biases" have been re-evaluated in this way (see e.g., Hilton, 1995).

This is not an attempt to find rationality at any cost; instead, it is *an interpretative strategy* that provides an essential methodological tool. This is further illustrated by ideal observer analysis as has been hugely successful in the study of perception (e.g., Geisler, 1987). Ideal observer models employ the formal tools of probability and decision theory to specify a model of optimal

[9]Again, the very fact that theories based on assumptions of rationality have come under increasing pressure within economics (both at the individual and the aggregate level, see e.g., Thaler and Mullainathan, 2008; Fox, 2010) is testimony to the fact that optimal models are falsifiable. At the same time, it is important to not confuse the fact that an empirical or theoretical assumption turns out to be wrong, or at some point needs to be replaced in order for a field to progress further, with the claim that greater insight and more rapid development would have been achieved without that assumption (cf. Elqayam and Evans, 2011; Jones and Love, 2011).

[10]That said, one may take issue with Elqayam and Evans construal of the relationship between is and ought in the context of explaining behavior more generally, see Corner and Hahn, 2013, for discussion.

performance given the available input for a task. Actual human performance is then compared to the performance of this ideal agent. In a process of iterative refinement, human performance and ideal observer are brought into closer and closer correspondence by incorporating capacity limitations of the human system into the ideal observer. This approach provides a tool for the *elucidation* of mechanism and process, embedded in an overall account that seeks to understand the system as "doing the best it can do" given the available hardware. In so doing, the approach inherently links behavioral prediction, mechanistic and functional explanation. In character, it might be viewed as a methodological formalization of the principle of charity.

Crucially, the aim is not to declare the system "optimal" *per se* (see also Griffiths et al., 2012 for related points on Bayesian modeling outside the context of ideal observer analysis). It remains the case that the (truly) optimal agent will be an ideal observer who is not subject to the many constraints of the human, physical system. So, to the extent that the human system achieves less than maximal performance, it is not "optimal" in the strongest possible sense, even if it is doing the best it can. At the same time, in the limit, a model that embodies *all* the constraints of the human system under scrutiny will just *be* that system. This means that, as a theoretical statement, it becomes increasingly vacuous to label a system as "optimal" (even in a weaker sense) as more and more constraints are built into the optimal agent to match its behavior (see also Jarvstad et al., 2014).

Instead, the point of the approach is a methodological one: rational models aide the disambiguation between competing theories and assist in the identification of underlying cognitive universals above and beyond the demand characteristics of experimental tasks (Howes et al., 2009). Once again, this gives such models and considerations a special status, above and beyond degrees of "model-fit" and so on.

SUMMARY AND CONCLUDING REMARKS

It has been argued in this paper that recent critiques of Bayesian modeling, and even more general critiques of computational level theories centered around normative considerations, are misdirected and misjudged. Specific examples have been used to counter any claim that Bayesian modeling would be inherently too flexible and thus unfalsifiable: not just the long-standing literature on judgment and decision-making, but more recent work within the context of reasoning and argumentation (e.g., Harris et al., 2012) provide ready examples of parameter-free model fits, where the model itself consists simply of Bayes' theorem.

It has also been claimed that for the development of "good" cognitive models other constraints (process level, or mechanism level) may be more important; against this, it has been highlighted that in many domains (in particular high-level domains such as reasoning or argumentation) the task participants face is one defined by inferential relationships between information states, and that an account that is based on those informational relationships is thus likely to explain most of the variance in behavioral prediction. That said, Bayesian accounts have been remarkably successful even in areas, such as perception (e.g., Knill and Richards, 1996; Yuille and Kersten, 2006), where mechanism can reasonably be expected to play a key role. Moreover, in many such domains, ideal observer analysis plays a valuable methodological role in identifying and understanding mechanistic constraints (Geisler, 1987). Hence the conflict between "mechanism" or "process" and rational explanation is methodologically ill-conceived. Pinning down processing constraints is likely to actually *require* appeal to optimality (see also, Howes et al., 2009).

At the same time, the present paper has given examples from within the reasoning and argumentation literature whereby Bayesian accounts, focussed on normative considerations, have demonstrably increased the level of behavioral prediction relative to that previously available in the relevant domain of research, and have provided analyses that open up (and first make possible) entirely new empirical programmes (a far cry from the accusation of merely recapitulating extant process/mechanism models). In all of this, this paper has sought to clarify why normative considerations (or considerations of "rationality" or "optimality") are theoretically interesting and methodologically important over and above behavioral prediction, potentially making a Bayesian model more than just another one of many competitors.

For any, or even all, of the examples used in setting out these arguments, the authors of the original critiques under scrutiny might wish to respond "but those are not the models I had in mind!." Certainly, Jones and Love (2011) claim only that Bayesian models frequently or maybe even typically exhibit some of the negative traits they perceive. Likewise, Bowers and Davis (2012a,b) supply a wealth of examples in making their case. The point of the present paper, however, is not to argue about whether or not certain perceptions are fair characterizations of the models that the authors of these critiques might have had in mind. Rather the point is to make the case that even if they were, the perceived limitations do not stem from the models *being Bayesian*. There could be a model or even many models for which some or all of the critiques examined here were apt and fair. However, the existence of examples to which the critiques do not apply indicates that it is not the formalism or Bayesian framework *per se* that would be to blame for any such inadequacy. Rather the fault would lie with the framework's particular application.

This matters because it constrains the debate about models. Whether typical or not, the examples chosen in this paper demonstrate that "Bayesian models" is *the wrong level of generality* at which to pitch these critiques. One may dislike specific models (or maybe even the models generally put forward by a specific modeler) and it will always be entirely proper to have debate about what supposedly makes a specific model "bad." But in order to best advance the quality of the models we as a discipline produce, such debate will need to be considerably more specific than the general critiques of Bayesian modeling examined here.

To some extent, all three of the critiques surveyed may be taken to agree with this, because each has sought to draw distinctions between types of Bayesian modeling [Jones and Love between "Bayesian Fundamentalism and Bayesian Enlightenment," Bowers and Davis between "Theoretical and Methodological Bayesianism," and Elqayam and Evans (2013) between "strict and soft Bayesianism"]. However, those distinctions themselves are motivated by the perceptions/claims that have been scrutinized in this paper. To the extent that these claims have been rejected, further classifications (and recommendations depending on them) are rejected also.

Cognitive modeling, however, does need more than debate about specific models. It arguably needs general debate about what exactly makes a model good, and the entire discipline arguably needs a better understanding of what, in general, makes explanation or theories "good" (for critiques of the state of psychological theorizing see e.g., Gigerenzer, 2009). It seems likely that the critiques by Jones and Love (2011), Bowers and Davis (2012a,b), and Elqayam and Evans (2011) evaluated here were motivated in part by disagreement about what aspects are most valuable in a cognitive model or theory. What those aspects should be and what kinds of theories and explanations we should strive for is a pressing issue. It is of great value if the critiques examined have started such debate.

ACKNOWLEDGMENT

The author was supported by the Swedish Research Council's Hesselgren Professorship.

REFERENCES

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychol. Rev.* 98, 409–429.

Armendt, B. (1993). Dutch books, additivity and utility theory. *Philol. Top.* 21, 1–20.

Bates, E., and MacWhinney, B. (1989). "Functionalism and the competition model," in *The Crosslinguistic Study of Sentence Processing*, eds B. MacWhinney and E. Bates (Cambridge: Cambridge University Press), 3–73.

Binmore, K. (1994). *Game Theory and the Social Contract, Volume I: Playing Fair*. Cambridge, MA: MIT Press.

Bowers, J. S., and Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* 138, 389–414. doi: 10.1037/a0026450

Bowers, J. S., and Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychol. Bull.* 138, 423–426. doi:10.1037/a0027750

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.* 78, 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brown, P. M. (1976). Conditionalization and expected utility. *Philos. Sci.* 43, 415–419. doi: 10.2307/187234

Busemeyer, J. R., and Bruza, P. D. (2012). *Quantum Models of Cognition and Decision*. Cambridge: Cambridge University Press.

Busemeyer, J. R., Pothos, E. M., Franco, R., and Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychol. Rev.* 118, 193. doi: 10.1037/a0022542

Camerer, C. (1995). "Individual decision making," in *Handbook of Experimental Economics*, eds J. Kagel and A. Roth (Princeton, NJ: Princeton University Press), 587–703.

Chater, N., and Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cogn. Psychol.* 38, 191–258.

Chater, N., Oaksford, M., Hahn, U., and Heit, E. (2010). Bayesian models of cognition. *WIREs Cogn. Sci.* 1, 811–823. doi: 10.1002/wcs.79

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: where next? *Trends Cogn. Sci.* 10, 335–344. doi: 10.1016/j.tics.2006.05.006

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use*. New York, NY: Praeger.

Corner, A., and Hahn, U. (2009). Evaluating scientific arguments: evidence, uncer-tainty & argument strength. *J. Exp. Psychol. Appl.* 15, 199–212. doi: 10.1037/a0016533

Corner, A. J., and Hahn, U. (2013). Normative theories of argumentation: are some norms better than others? *Synthese* 190, 3579–3610. doi: 10.1007/s11229-012-0211-y

Corner, A. J., Hahn, U., and Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *J. Mem. Lang.* 64, 153–170. doi: 10.1016/j.jml.2010.10.002

Cox, G. W. (1999). The empirical content of rational choice theory: a reply to Green and Shapiro. *J. Theor. Polit.* 11, 147–169.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. New York, NY: Springer.

Earman, J. (1992). *Bayes or Bust?* Cambridge, MA: MIT Press.

Edgington, D. (1995). On conditionals. *Mind* 104, 235–329.

Edwards, W. (1968). "Conservatism in human information processing," in *Formal Representation of Human Judgment,* ed B. Kleinmuntz (New York, NY: Wiley), 17–52.

Elqayam, S., and Evans, J. S. B. T. (2011). Subtracting "ought" from 'is': descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Elqayam, S., and Evans, J. S. B. T. (2013). Rationality in the new paradigm: strict versus soft Bayesian approaches. *Think. Reason.* 19, 453–470. doi: 10.1080/13546783.2013.834268

Evans, J. S. B. T., and Over, D. E. (2004). *If.* Oxford: Oxford University Press.

Ferejohn, J. A. (2002). Symposium on explanations and social ontology 1: rational choice theory and social explanation. *Econ. Philos.* 18, 211–234. doi: 10.1017/S026626710200202X

Fodor, J. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fox, J. (2010). *The Myth of the Rational Market: A History of Risk, Reward, and Delusion on Wall Street*. Petersfield: Harriman House Publishing.

Friedman, M., and Savage, L. (1948). 'The utility analysis of choices involving risk.' *J. Polit. Econ*. LVI, 279–304.

Friedman, M., and Savage, L. (1952). 'The expected-utility hypothesis and the measurability of utility.' *J. Polit. Econ*. LX, 463–74.

Geisler, W. S. (1987). "Ideal-observer analysis of visual discrimination," in *Frontiers of Visual Science: Proceedings of the 1985 Symposium (Committee on Vision ed),* (Washington, DC: National Academy Press), 17–31. Available online at: http://searchworks.stanford.edu/view/1295674

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741.

Gigerenzer, G. (2009). Surrogates for theory. *Observer* 22, 21–23. Available online at: http://www.psychologicalscience.org/index.php/publications/observer/2009/february-09/surrogates-for-theory.html

Gold, E. (1967). Language identification in the limit. *Inf. Control* 16, 447–474.

Govier, T. (1987). *Problems in Argument Analysis and Evaluation*. Dordrecht: Foris Publications.

Griffiths, T. L., Chater, N., Norris, D., and Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012). *Psychol. Bull.* 138, 415–422. doi: 10.1037/a0026884

Griffiths, T. L., and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cogn. Psychol.* 51, 334–384. doi: 10.1016/j.cogpsych.2005.05.004

Hahn, U. (2009). Explaining more by drawing on less. Commentary on Oaksford, M. & Chater, N. *Behav. Brain Sci.* 32, 90–91. doi: 10.1017/S0140525X09000351

Hahn, U. (2011). Why rational norms are indispensible. Commentary on Elqayam and Evans. *Behav. Brain Sci.* 34, 257–258. doi: 10.1017/S0140525X11000641

Hahn, U., and Chater, N. (1997). "Concepts and similarity," in *Knowledge, Concepts and Categories,* eds K. Lamberts and D. Shanks (Hove: Psychology Press: MIT Press), 43–92

Hahn, U., Harris, A. J., and Corner, A. (2009). Argument content and argument source: an exploration. *Informal Logic* 29, 337–367.

Hahn, U., and Oaksford, M. (2007). The rationality of informal argumentation: a Bayesian approach to reasoning fallacies. *Psychol. Rev.* 114, 704–732. doi: 10.1037/0033-295X.114.3.704

Hahn, U., and Oaksford, M. (2008). "Inference from absence in language and thought," in *The Probabilistic Mind,* eds N. Chater and M. Oaksford (Oxford: Oxford University Press), 121–142.

Hahn, U., and Oaksford, M. (2012). "Rational argument," in *Oxford Handbook of Thinking and Reasoning*, eds R. Morrison and K. Holyoak (Oxford: Oxford University Press), 277–298.

Hajek, A. (2008). "Dutch book arguments," in *The Handbook of Rational and Social Choice,* eds P. Anand, P. Pattanaik, and C. Puppe (Oxford: Oxford University Press), 173–196.

Harris, A. J. L., and Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: incorporating the role of coherence. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1366–1373. doi: 10.1037/a0016567

Harris, A. J. L., Hsu, A. S., and Madsen, J. K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using ad hominem. *Think. Reason.* 18, 311–343. doi: 10.1080/13546783.2012.670753

Hilton, D. (1995). The social context of reasoning: conversational inference and rational judgment. *Psychol. Bull.* 118, 248–271. doi: 10.1037/0033-2909.118.2.248

Howes, A., Lewis, R. L., and Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychol. Rev.* 116, 717–751. doi: 10.1037/a0017187

Howson, C., and Urbach, P. (1993). *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.

Jarvstad, A., Hahn, U.,Warren, P., and Rushton, S. (2014). Are perceptuo-motor decisions really more optimal than cognitive decisions? *Cognition* 130, 397–416. doi: 10.1016/j.cognition.2013.09.009

Jones, M., and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav. Brain Sci.* 34, 169–188. doi: 10.1017/S0140525X10003134

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philos. Sci.* 65, 573–603.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185

Karni, E. (2014). "Axiomatic foundations of expected utility and subjective probability," in *Handbook of the Economics of Risk and Uncertainty*, Vol. 1, eds J. Mark, W. Machina, and K. Viscusi (Oxford: North Holland).

Kemp, C., and Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychol. Rev.* 116, 20–57. doi: 10.1037/a0014282

Klayman, J., and Ha, Y. (1989). Confirmation, disconfirmation, and information in hypothesis testing. *Psychol. Rev.* 94, 211–228. doi: 10.1037/0033-295X.94.2.211

Klauer, K. C., Stahl, C., and Erdfelder, E. (2007). The abstract selection task: new data and an almost comprehensive model. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 680–703. doi: 10.1037/0278-7393.33.4.680

Knill, D. C., and Richards, W. (eds.). (1996). *Perception as Bayesian inference*. Cambridge: University Press.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Lakatos, I. (1976). "Falsification and the methodology of scientific research programmes," in *Can Theories be Refuted*? ed S. G. Harding (Netherlands: Springer), 205–259. doi: 10.1007/978-94-010-1863-0_14

Lehtinen, A., and Kuorikoski, J. (2007). Unrealistic assumptions in rational choice theory. *Philos. Soc. Sci.* 37, 115–138. doi: 10.1177/0048393107299684

Leitgeb, H., and Pettigrew, R. (2010a). An objective justification of bayesianismi: measuring inaccuracy*. *Philos. Sci.* 77, 201–235. doi: 10.1086/651317

Leitgeb, H., and Pettigrew, R. (2010b). An objective justification of bayesianism ii: the consequences of minimizing inaccuracy*. *Philos. Sci.* 77, 236–272. doi: 10.1086/651318

Lewandowsky, S., Gignac, G. E., and Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nat. Clim. Change* 3, 399–404. doi: 10.1038/nclimate1720

Neuman, Y., and Weitzman, E. (2003). The role of text representation in students' ability to identify fallacious arguments. *Q. J. Exp. Psychol.* 56A, 849–864. doi: 10.1080/02724980244000666

Newell, B. R., Lagnado, D. A., and Shanks, D. R. (2007). *Straight Choices: The Psychology of Decision Making*. Hove: Psychology Press.

Oaksford, M. (2014). Normativity, interpretation, and Bayesian models. *Front. psychol.* 5:332. doi: 10.3389/fpsyg.2014.00332

Oaksford, M., and Chater, N. (1991). Against logicist cognitive science. *Mind Lang.* 6, 1–38.

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 608–631.

Oaksford, M., and Chater, N. (2009). Précis of Bayesian rationality: the probabilistic approach to human reasoning. *Behav. Brain Sci.* 32, 69–84. doi: 10.1017/S0140525X09000284

Oberauer, K. (2006). Reasoning with conditionals: a test of formal models of four theories. *Cogn. Psychol.* 53, 238–283. doi: 10.1016/j.cogpsych.2006.04.001

Perfors, A., Tenenbaum, J. B., and Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition* 118, 306–338. doi: 10.1016/j.cognition.2010.11.001

Peterson, C. R., and Beach, L. R. (1967). Man as an intuitive statistician. *Psychol. Bull.* 68, 29–46.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson & Co.

Pothos, E. M., and Busemeyer, J. R. (2009). A quantum probability explanation for violations of 'rational' decision theory. *Proc. Biol. Sci.* 276, 2171–2178. doi: 10.1098/rspb.2009.0121

Pothos, E. M., and Busemeyer, J. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behav. Brain Sci.* 36, 255–327. doi: 10.1017/S0140525X12001525

Pothos, E. M., and Busemeyer, J. (2014). In search for a standard of rationality. *Front. Psychol.* 5:49. doi: 10.3389/fpsyg.2014.00049

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., et al. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition* 121, 83–100. doi: 10.1016/j.cognition.2011.06.002

Rabin, M., and Thaler, R. H. (2001). Anomalies: risk aversion. *J. Econ. Perspect.* 59, 219–232. doi: 10.1257/jep.15.1.219

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rosenkrantz, R. D. (1992). The justification of induction. *Philos. Sci.* 15, 527–539. doi: 10.1086/289693

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.* 117, 1144–1167. doi: 10.1037/a0020511

Satz, D., and Ferejohn, J. A. (1994). Rational choice and social theory. *J. Philos.* 91, 71–87.

Savage, L. J. (1954). *The Foundations of Statistics*. New York, NY: Wiley.

Schroyens, W., and Schaeken, W. A. (2003). A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 140–149. doi: 10.1037/0278-7393.29.1.140

Seidenberg, M. S., and MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cogn. Sci.* 23, 565–588.

Skyrms, B. (1993). A mistake in dynamic coherence arguments? *Philos. Sci.* 60, 320–328.

Starmer, C. (2005). Normative notions in descriptive dialogues. *J. Econ. Methodol.* 12, 277–289. doi: 10.1080/13501780500086206

Teller, P. (1973). Conditionalization and observation. *Synthese* 26, 218–258.

Tentori, K., and Crupi, V. (2013). Why quantum probability does not explain the conjunction fallacy. *Behav. Brain Sci.* 36, 308–310. doi: 10.1017/S0140525X12003123

Thaler, R. H., and Mullainathan, S. (2008). "*Behavioral Economics*," *The Concise Encyclopedia of Economics, 2nd Edn*. Indianapolis, IN: Liberty Fund.

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315.

Von Neumann, J., and Morgenstern, O. (1947). *The Theory of Games and Economic Behaviour, 2nd Edn*. Princeton, NJ: Princeton University Press.

Wakker, P. P. (2010). *Prospect Theory for Risk and Ambiguity*. Cambridge: Cambridge University Press.

Wason, P. (1968). Reasoning about a rule. *Q. J. Exp. Psychol.* 20, 273–281.

Wharton, R. (1974). Approximate language identification. *Inf. Control* 26, 236–255.

Winkler, R. L., and Murphy, A. H. (1968). "Good" probability assessors. *J. Appl. Meteorol.* 7, 751–758.

Yuille, A., and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis?. *Trends Cogn. Sci.* 10, 301–308. doi: 10.1016/j.tics.2006.05.002

Conflict of Interest Statement: The Reviewer, Mike Oaksford, declares that despite being affiliated to the same institution as the author Ulrike Hahn, the review process was handled objectively and no conflict of interest exists. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 March 2014; paper pending published: 30 April 2014; accepted: 30 June 2014; published online: 08 August 2014.

*Citation: Hahn U (2014) The Bayesian boom: good thing or bad?. Front. Psychol. **5**:765. doi: 10.3389/fpsyg.2014.00765*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

OPINION ARTICLE
published: 13 March 2014
doi: 10.3389/fpsyg.2014.00219

From is to ought, and back: how normative concerns foster progress in reasoning research

Vincenzo Crupi[1]* and Vittorio Girotto[2]

[1] *Department of Philosophy and Education, University of Turin, Turin, Italy*
[2] *Center for Experimental Research in Management and Economics, University IUAV of Venice, Venice, Italy*
**Correspondence: vincenzo.crupi@unito.it*

Edited by:
David E. Over, Durham University, UK

Reviewed by:
Shira Elqayam, De Montfort University, UK

Keywords: rationality, reasoning, selection task, pseudodiagnosticity, conjunction fallacy

INTRODUCTION

Can the issue of human (ir)rationality contribute to the scientific study of reasoning? A tempting line of argument seems to indicate that it can't. Here it is. (i) To discuss diagnoses of (ir)rationality arising from research in the psychology of reasoning one has to deal with *arbitration*, i.e., the assessment of competing theories of what a reasoner ought to do, if rational. But (ii), by the Humean divide between *is* and *ought*, arbitration is logically independent from the description of reasoning. And clearly (iii) the main goal of psychological inquiry is just such a description. It follows that normative concerns about diagnoses of (ir)rationality cannot serve the proper scientific purposes of the psychology of reasoning, and would better be left aside altogether in this area. A recent cornerstone for this debate is Elqayam and Evans (2011). Part of their discussion is devoted to voice precisely this criticism of "normativism," thus favoring a purely "descriptivist" approach in the study of human thinking. In our view, the above argument is essentially valid, but unsound. Premise (i), in particular, may have seemed obvious but doesn't hold on closer inspection, as we mean to show.

In reasoning experiments, participants are assumed to rely on some amount of information, or data, D. These include elements explicitly provided (e.g., a cover story), but possibly also further background assumptions. Note that, as a rule, D is *not* already framed in a technical language such as that of, say, probability theory: cover stories and experimental scenarios are predominantly verbal in nature, although they may embed more formal fragments (e.g., some statistical information). On the basis of D, participants then have to produce one among a set of possible responses R, for instance an item chosen in a set of options or an estimate in a range of values allowed (say, 0 to 100%). Here again, the possible responses do *not* belong to a particular formal jargon (although, again, some formal bits may occur in the elements of R).

Suppose that some particular response r in R turns out to be widespread among human reasoners and is said to be irrational. Such a diagnosis, we submit, has to rely on *four* premises. (i) First, one has to identify a formal theory of reasoning T as having normative force[1]. (ii) Second, one has to map D onto a formalized counterpart D^* belonging to the technical language employed in T. (iii) Third, one has to map R, too, onto a formalized counterpart R^* belonging to the technical language of T. This step implies, in particular, that the target response r within R be translated into its appropriate counterpart r^*. (iv) And finally, one has to show that, given D^*, r^* does contradict T. If either of (i)–(iv) is rejected, the charge of irrationality fails. We thus have a classification of the ways in which one can question diagnoses of irrationality that may be attached to the results of a reasoning experiment. Depending on whether (i), (ii), (iii), or (iv) is the main focus of controversy, we will talk about *arbitration*, *data mismatch*, *response mismatch*, and *norm misapplication*, respectively. Relying on this partition, let us now consider three prominent cases in which normative concerns have entered psychological research on reasoning.

[1] We emphasize that here we are not committed in any way to the idea of T as a "computational model" or a "theory of competence," as they are often understood. Such a move would risk to blur our current analysis (we concur with Evans and Elqayam, 2011: 277, and others on at least this much). Of course, T will be a formal system—say, classical probability theory. But, according to (i), in order for a diagnosis of irrationality to hold, T has to be taken as having normative force, namely, with an additional overarching claim that a rational agent ought to comply with its principles.

EXHIBIT 1: THE SELECTION TASK AND DATA MISMATCH

The debate on Wason's selection task is said to have sparked the rise of a new paradigm in the psychology of reasoning (see, e.g., Over, 2009), and so it seems a primary example of how progress in this field can intertwine with diverging diagnoses of rational behavior (see Sperber et al., 1995, though, for cautionary considerations). In the standard version of the selection task, four cards are employed which have a letter on one side and a number on the other side. One can see the letter-side of two cards (A and C, say), and the number-side of the other two (4 and 7, say). Which of these cards would one need to turn over to decide whether the following statement is true or false? "If there is a vowel on the one side, then there is an even number on the other side." In the classical analysis of the selection task, this statement was interpreted as a material conditional and referred to the four cards only. The statement would then be true unless some of the four cards has a vowel and an odd number. Accordingly, the A and the 7 cards ought to be turned over; the C and the 4 cards are of no use,

logically. Participants often selected the 4 card, largely disregarding the 7 card, and were thus charged of being irrational.

In Oaksford and Chater's (1994, 2003) work, however, the ordinary language sentence "if vowel, then even number" is not taken as a material conditional, but rather as such that its probability is the conditional probability that the card has an even number on one side given that it has a vowel on the other side. Moreover, the conditional statement is referred to a larger deck of which the four cards only represent a sample and in which, finally, the occurrence of both vowels and even numbers are assumed to be relatively rare. This radically different formal reconstruction of the data D defining the problem has important consequences. The implication that, for instance, turning over a card showing number 4 is irrational does not hold anymore and an alternative normative analysis is required (see Fitelson and Hawthorne, 2010). In our current terms, the key point of this debate is a matter of *data mismatch*. Importantly, no doubt needs to be raised against the normative status of classical logic to make sense of this case. (A parallel account could be given for non-probabilistic approaches such as Stenning and van Lambalgen's 2008).

EXHIBIT 2: THE CONJUNCTION FALLACY AND RESPONSE MISMATCH

Upon experimental investigation, individuals often rank a conjunctive statement "x and y" as more probable than one of the conjuncts (e.g., x). For instance, most physicians judge that a patient who had pulmonary embolism is more likely to experience "dyspnea and hemiparesis" than "hemiparesis." Tversky and Kahneman (1983) famously labeled this a fallacy, because in probability theory $Pr(x \wedge y) = Pr(x)$ for any x, y, regardless of what information may be available. Note that the latter clause prevents rescue of the rationality of human judgment by an appeal to data mismatch. In fact, in debates about the conjunction fallacy, it is *response mismatch* that has been relentlessly discussed. Given how fundamental and startling this judgment bias seemed, almost all conceivable worries have been aired over the years. Maybe, in the presence of a conjunctive statement "x and y," pragmatic considerations led participants to treat the isolated conjunct "x" as "$x \wedge$ not-y." Or maybe the ordinary language conjunction "x and y" was mapped onto a logical disjunction ("$x \vee y$"), or a conditional expression ("y, assuming that x"). Or the quantities to be ranked were not meant to be $Pr(x \wedge y)$ and $Pr(x)$ because the reference of the ordinary language term "probable" eluded the basic properties of mathematical probability. In each of these cases, the suggested rendition r^* of the modal response r (here: that statement "x and y" was more probable than "x") would have *not* contradicted probability theory, thus deflating the charge of irrationality.

Here again, there is no logical reason to saddle this debate with any subtlety concerning the normative appeal of the target formal theory (classical probability) for human reasoning. And while all of the above worries of response mismatch had been already addressed by Tversky and Kahneman (1983) (see, e.g., Girotto, 2011), their recurrent appearance in the literature spurred the development of more and more refined experimental techniques leading to a better understanding of this reasoning bias. (See Wedell and Moro, 2008; Tentori and Crupi, 2012, 2013; Tentori et al., 2013 for discussions).

EXHIBIT 3: PSEUDODIAGNOSTICITY AND NORM MISAPPLICATION

In its simplest form (e.g., Kern and Doherty, 1982), so-called pseudodiagnosticity task provides participants with a binary set of blank and equiprobable hypotheses h and $\neg h$ (e.g., two abstract diagnoses), two pieces of evidence e and f (e.g., two symptoms) and one likelihood value, such as $Pr(e|h) = 65\%$. Participants have to select the most useful among three further likelihood values, $Pr(e|\neg h)$, $Pr(f|h)$, and $Pr(f|\neg h)$. In the classical interpretation of this phenomenon, participants were said to have "actively chose[n] irrelevant information [namely, $Pr(f|h)$] and ignored relevant information [namely, $Pr(e|\neg h)$] which was equally easily available" (Doherty et al., 1979, p. 119). The standard Bayesian framework was taken as a benchmark theory sanctioning this conclusion. But the idea of so-called pseudodiagnosticity bias was seen by Crupi et al. (2009) as a case of *norm misapplication*.

Crupi et al. (2009) offered formal renditions (D^* and R^*, in our notation) of the experimental scenario (D) and the response set (R) that were consistent with the classical reading of the task (so they argued on the basis of textual evidence). Thus no data or response mismatch was invoked, in our current terms. Crupi et al., submitted, instead, that the relevant norms of reasoning had been misapplied in the standard interpretation: far from contradicting the benchmark theory, the appropriate formal counterpart r^* of the participants' modal response r in pseudodiagnosticity experiments turns out to be actually optimal for a Bayesian agent (given D^*). Tweeney et al. (2010), in turn, criticized this conclusion. However, they outlined themselves a further novel theoretical analysis of the task and did *not* try to revive the once popular interpretation of the phenomenon in its original form. To the extent that the latter is now judged inadequate by all parties involved, at least some theoretical progress was made whatever the outcome of this debate.

CONCLUDING REMARKS

According to a seductive argument, debates on the (ir)rationality of participants' responses are better left out of the psychologist's outlook for they would invariably led her to plod on the shaky ground of arbitration. We have challenged this assumption by means of three key examples. The selection task, the conjunction fallacy and pseudodiagnosticity have been extensively investigated in the psychology of reasoning, and all raised lively controversies about human rationality. Yet, issues of arbitration hardly played any substantive role. Once properly reconstructed, the relevant problem was not whether it is rational to depart from the implications of allegedly compelling normative theories such as logic or probability theory. Instead, much of the research done with these classical paradigms was focussed on whether and how those implications could connect with observed behavior given that data mismatch, response mismatch or norm misapplication may have occurred.

Arbitration between competing norms of reasoning is central to certain areas of

philosophy but remains marginal in psychological research, and for good reasons, loosely related to the so-called *is*/*ought* divide: arbitration does require specific forms of argumentation that lie outside the usual scope of empirical research (see, e.g., Schurz, 2011; Pettigrew, 2013). Concerns of data mismatch, response mismatch and norm misapplication, on the contrary, are amenable to independent scrutiny in purely descriptive terms (be that at the empirical or theoretical level). Sometimes earlier charges of irrationality and biased reasoning survived increasingly stringent demands of this kind (the conjunction fallacy is a case in point), sometimes not (pseudodiagnosticity illustrates). Either way, a significant amount of theoretical and/or experimental insight has been achieved. We conclude that normative concerns about diagnoses of (ir)rationality can retain a legitimate and constructive role for the psychology of reasoning.

ACKNOWLEDGMENTS

Vincenzo Crupi acknowledges support from the Italian Ministry of Scientific Research (FIRB project *Structures and Dynamics of Knowledge and Cognition*, Turin unit, D11J12000470001) and from the Deutsche Forshungsgemeinschaft (priority program *New Frameworks of Rationality*, SPP 1516, grant CR 409/1-1). Vittorio Girotto acknowledges support from the Swiss and Global—Ca' Foscari Foundation and from the Italian Ministry of Scientific Research (PRIN grant 2010-RP5RNM).

REFERENCES

Crupi, V., Tentori, K., and Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychol. Rev.* 116, 971–985. doi: 10.1037/a0017050

Doherty, M. E., Mynatt, C. R., Tweeney, R. D., and Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychol.* 43, 111–121. doi: 10.1016/0001-6918(79)90017-9

Elqayam, S., and Evans, J. S. (2011). Subtracting "ought" from "is": Descriptivism *versus* normativism is the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Evans, J. S., and Elqayam, S. (2011). Towards a descriptivist psychology of reasoning and decision making. *Behav. Brain Sci.* 34, 275–284. doi: 10.1017/S0140525X11001440

Fitelson, B., and Hawthorne, J. (2010). The Wason task(s) and the paradox of confirmation. *Philos. Perspect.* 24, 207–241. doi: 10.1111/j.1520-8583.2010.00191.x

Girotto, V. (2011). Undisputed norms and normal errors in human thinking. *Behav. Brain Sci.* 34, 255–256. doi: 10.1017/S0140525X11000483

Kern, L., and Doherty, M. E. (1982). Pseudodiagnosticity' in an idealized medical problem-solving environment. *J. Med. Educ.* 57, 100–104.

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 608–631. doi: 10.1037/0033-295X.101.4.608

Oaksford, M., and Chater, N. (2003). Optimal data selection: revision, review, and re-evaluation. *Psychon. Bull. Rev.* 10, 289–318. doi: 10.3758/BF03196492

Over, D. E. (2009). New paradigm psychology of reasoning. *Think. Reason.* 15, 431–438. doi: 10.1080/13546780903266188

Pettigrew, R. (2013). Epistemic utility and norms for credences. *Philos. Comp.* 8, 897–908. doi: 10.1111/phc3.12079

Schurz, G. (2011). Truth-conduciveness as the primary epistemic justification of normative systems of reasoning. *Behav. Brain Sci.* 34, 266–267. doi: 10.1017/S0140525X11000537

Sperber, D., Cara, D., and Girotto, V. (1995). Relevance theory explains the selection task. *Cognition* 57, 31–95. doi: 10.1016/0010-0277(95)00666-M

Stenning, K., and van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science.* Cambridge: MA: MIT Press.

Tentori, K., and Crupi, V. (2012). On the conjunction fallacy and the meaning of *and*, yet again: a reply to Hertwig, Benz, and Krauss (2008). *Cognition* 122, 123–134. doi: 10.1016/j.cognition.2011.09.002

Tentori, K., and Crupi, V. (2013). Why quantum probability does not explain the conjunction fallacy. *Behav. Brain Sci.* 36, 308–310. doi: 10.1017/S0140525X12003123

Tentori, K., Crupi, V., and Russo, S. (2013). On the determinants of the conjunction fallacy: probability *versus* inductive confirmation. *J. Exp. Psychol. Gen.* 142, 235–255. doi: 10.1037/a0028770

Tversky, A., and Kahneman, D. (1983). Extensional *versus* intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Tweeney, R. D., Doherty, M. E., and Kleiter, G. D. (2010). The pseudodiagnosticity trap: should participants consider alternative hypotheses? *Think. Reason.* 16, 332–345. doi: 10.1080/13546783.2010.525860

Wedell, D. H., and Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: effects of response mode, conceptual focus, and problem type. *Cognition* 107, 105–136. doi: 10.1016/j.cognition.2007.08.003

Received: 29 November 2013; accepted: 25 February 2014; published online: 13 March 2014.

Citation: Crupi V and Girotto V (2014) From is to ought, and back: how normative concerns foster progress in reasoning research. Front. Psychol. ***5****:219. doi: 10.3389/fpsyg.2014.00219*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 27 May 2014
doi: 10.3389/fpsyg.2014.00503

How (not) to argue about is/ought inferences in the cognitive sciences

Katinka J. P. Quintelier[1]* and Lieuwe Zijlstra[2]

[1] *Department of International Strategy and Marketing, Amsterdam Business School, University of Amsterdam, Amsterdam, Netherlands*
[2] *Department of Philosophy and Moral Sciences, Ghent University, Ghent, Belgium*
**Correspondence: k.quintelier@uva.nl*

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Gerhard Schurz, Heinrich Heine University Duesseldorf, Germany

Keywords: is/ought gap, naturalistic fallacy, is/ought inferences, epistemic "oughts", deontic "oughts", defeasible reasoning, deontic reasoning

When scholars problematize is/ought inferences (IOI's), they sometimes refer to Hume's or Moore's fallacy (e.g., Schneider, 2000; Schroyens, 2009; Elqayam and Evans, 2011). Although inferring "ought" from "is" can be problematic, we argue that, in the context of contemporary IOI's in the cognitive sciences, invoking Hume or Moore might be misguided. This is because Hume's and Moore's arguments concern the validity and soundness of *deductive* inferences while in our view contemporary IOI's in the cognitive sciences are better interpreted as *defeasible* inferences.

In order to avoid misinterpretations, we first clarify key concepts in the debate in section Key Concepts. In section Mind the Gap, we revisit Hume's and Moore's arguments against inferring "ought" from "is," and in section A Debate Shackled, we discuss contemporary IOI's in the cognitive sciences.

KEY CONCEPTS

Participants in the is/ought debate distinguish between descriptive statements and deontic statements. Descriptive statements describe or predict how the world is. Deontic statements prescribe or proscribe how we should act or reason.

While "is" statements are descriptive statements, "ought" statements can be descriptive as well as deontic. For instance, "the streets ought to be wet because it is raining" is a descriptive statement because it predicts that the streets will be wet. Conversely, "If you do not want to get wet, you ought to carry an umbrella," is a deontic statement because it prescribes what you should do. In this comment, we only discuss "ought" statements as deontic statements. Accordingly, we will not discuss inferences from "is" to descriptive "oughts" (cf. Oaksford and Chater, 2009, 2011), but only inferences from "is" to deontic "oughts" (cf. Oaksford and Sellen, 2000; Stanovich and West, 2000).

We describe an is/ought inference as an attempt to evaluate (i.e., fine-tune, develop, arbitrate between) deontic statements on the basis of descriptive statements. The following is an example of an IOI:

(1) Premise: More intelligent people are more likely than less intelligent people to make a guess, instead of reason, when solving the Wason Selection Task.
Conclusion: We ought to make a guess, instead of reason, when solving the Wason Selection Task.

This inference can be interpreted as a deductive argument. As such, the conclusion is true if the inference is valid and sound. A deductive inference is valid if the premises logically entail the conclusions, hence, if it is logically impossible for the premises to be true and the conclusion false. In this inference, it is possible that the premise is true while the conclusion is false. Thus, it is deductively invalid.

Soundness takes the actual truth of the premises (and conclusions) into account: An inference is sound if it is valid *and* all of its premises are true. The inference in this example is not sound because it is invalid. However, were it to be valid, it would still be unsound because the premise is false. More intelligent people are in fact more likely than less intelligent people to reason logically when solving the Wason Selection Task (Stanovich and West, 2000).

An inference can also be interpreted as a defeasible argument. Defeasible inferences have several features, two of which are relevant for our argument (cf. Pollock, 1987, 1992). First, the inference can be correct even if it is not deductively valid. Let us illustrate this features on the basis of the following inference (which is not an is/ought inference) (2):

(2) Premise: X looks red to me.
Conclusion: X is red.

Clearly, the premise does not logically entail the conclusion. However, the inference is defeasibly correct because the premise supports the conclusion—most things that look red to me are, in fact, red.

A second feature of defeasible inferences is that, when the inference is correct, it can still be revised in the light of new information. For instance, if we learn that X is a daisy that is illuminated by red lights, which can make things appear red when they are not, we may suggest the following revised inference (3):

(3) Premise 1: X looks red to me.
Premise 2: X is a daisy that is illuminated by red lights, which can make things appear red when they are not.
Conclusion: X is not red.

While correct defeasible inferences can be revised in the light of new information, valid deductive inferences cannot: If the conclusion follows deductively from a (set of) premise(s), it will still follow deductively no matter how many premises we add. (This is termed the monotonicity of deductive logic.)

All this is relevant for is/ought debates. In section Mind the Gap, we argue that Hume's and Moore's arguments concern the validity and soundness of deductive inferences. In section A Debate Shackled, we explain why IOI's in the cognitive sciences are better interpreted—and evaluated—as defeasible inferences.

MIND THE GAP

Cognitive scientists often fine-tune, develop or arbitrate between models of how people ought to reason on the basis of theories and data of how people do reason (for a discussion and critique, see Elqayam and Evans, 2011). Critics (e.g., Schneider, 2000; Schroyens, 2009; Elqayam and Evans, 2011) claim that some of these cognitive scientists commit Hume's or Moore's fallacy. However, in line with previous interpretations, we contend that Hume's and Moore's fallacies in the first place preclude deductive inferences that are, respectively, not valid and not sound (cf. Schurz, 1997; Pigden, 2010; Quintelier et al., 2011).

It is useful to introduce a caveat here. Hume and Moore formulated their arguments in the context of ethical "oughts." However, in the cognitive sciences, their arguments are applied to epistemic "oughts." This is acceptable for standard, logical, interpretations of Hume's fallacy, which seem to hold at least for deontic "oughts" in general (Pigden, 2010; P. 240). In contrast, it is unclear if Moore's fallacy applies to the same extent to non-ethical deontic "oughts." For the sake of argument though, we assume that both fallacies also apply to epistemic "oughts."

Let us now review Hume's fallacy. The standard interpretation of Hume's fallacy states that there are no deductively valid inferences whose premises contain no "oughts" and whose conclusions contain (non-trivial) "oughts" (Schurz, 1997; Pigden, 2010; p. 198–242). For example, the following inference is not deductively valid:

(3) Premise: It is the case that human beings apply Bayesian reasoning.
Conclusion: It ought to be the case that human beings apply Bayesian reasoning.

This inference is not deductively valid because it is possible that the conclusions are false while the premises are true. In Hume's words, "ought, or ought not, expresses some new relation or affirmation," which is different than the relation being expressed by "is," or "is not" (1739–1740, Book III, Part I, section Key Concepts). When scholars infer "ought" related conclusions from premises that contain only "isses," they commit Hume's fallacy.

However, Hume also argues that we can add a premise—hereafter termed a bridge principle - that connects "is" and "ought." We can for example suggest the following bridge principle: "if more intelligent people apply reasoning X, we ought to apply reasoning X" (cf. Schneider, 2000, commenting on Stanovich and West, 2000). This principle can then be used as a premise:

(4) Premise 1: More intelligent people apply Bayesian reasoning.
Premise 2: If more intelligent people apply Bayesian reasoning, we ought to apply Bayesian reasoning.
Conclusion: We ought to apply Bayesian reasoning.

This inference is now deductively valid: if the premises are true, then the conclusion is also true. Hume's fallacy does not preclude the possibility of finding a plausible bridge principle.

In contrast, Moore's fallacy states that deductive IOI's with bridge principles might be valid, but they are never sound. The reason is that, according to Moore, bridge principles can never be true. Moore's argument is that we should find an analytically true bridge principle, one that spells out what descriptive concepts are in the meaning of the deontic concept (Moore, 1988, §1–15). However, pace Moore, this is impossible because deontic concepts are already simple terms; there is nothing in their meaning than the deontic concept itself. Therefore, there are no true bridge principles. Those who define a deontic concept in descriptive terms and then claim that this definition is analytically true, commit Moore's fallacy (id.).

To summarize, we hold that Hume's fallacy states that deductive IOI's are never valid without a bridge principle, while Moore's thesis states that deductive IOI's are never sound because there is no true bridge principle.

A DEBATE SHACKLED

Invoking Hume's and Moore's fallacy to criticize IOI's in the cognitive sciences can be problematic: If, by making an is/ought inference, authors rarely mean to *deduce* deontic "oughts" from "isses," then their IOI's should not be evaluated on the basis of their deductive validity or soundness. Indeed, we argue that it is more charitable to interpret contemporary IOI's in the cognitive sciences as defeasible inferences: Relevant authors (Oaksford and Sellen, 2000; Stanovich and West, 2000; Douven, 2011) point to descriptive reasons that suggest, rather than logically entail, deontic conclusions. Moreover, these authors aim to make correct inferences that are revisable in the light of new information. Let us take a look at these features of contemporary IOI's in the cognitive sciences.

Stanovich and West (2000) seem to endorse the following inference:

(5) Premise: Studies show that more intelligent people are more likely than less intelligent people to reason logically in task A.
Conclusion: We ought to reason logically in task A.

Oaksford and Sellen (2000) remark that the following also holds:

(6) Premise: Studies show that high schizotypal people are more likely than low schizotypal people to reason logically in task B.
Conclusion: We ought not to reason logically in task B.

Clearly, these inferences are not deductively valid (cf. Schneider, 2000). However, these authors never claimed that their premise deductively entails a deontic conclusion. Instead, both Stanovich and West (2000; p. 645) and Oaksford and Sellen (2000; p. 691) speak of descriptive information that *suggests* a certain deontic

conclusion. Moreover, these arguments are revisable in the light of new information: What if, for instance, both schizotypy and intelligence are positively correlated with logical reasoning in the same task A? In that case, we have to revise our conclusions that we ought to reason logically in task A. Thus, inferences 5 and 6 are better understood as defeasible inferences and ought to be evaluated accordingly.

Douven (2011) likewise suggests that, in certain cases, descriptive information can be used to inform us about deontic statements. He reasons as follows:

(7) Premise: Human beings update on conditionals by applying rule X.
Conclusion: Human beings ought to update on conditionals by applying rule X.

Again, as a deductive inference, this would be invalid. However, Douven (2011) does not seem to have a deductive inference in mind. In his words, the premise again "suggests" the conclusion, and descriptive information leads to an "outline" of norms or, based on the premise, we can go "some way" in accepting the conclusion (253). This can be understood as a first approximation that can be revised. Moreover, there is no mentioning that descriptive premises logically entail a deontic conclusion.

These examples lead us to conclude that IOI's in the cognitive sciences are better interpreted as defeasible inferences than as deductive inferences. As a consequence, their deductive validity and soundness is not at stake. We therefore suggest that, instead of referring to Hume or Moore, critics of is/ought inferences apply evaluation criteria for defeasible inferences (see e.g., Nute, 1997). This conclusion supplements previous work on the is/ought problem. Schurz (in Pigden, 2010; p. 216), for instance, suggests that defeasible conditional norms might provide plausible bridge principles in ethical is/ought inferences. Other authors suggest that defeasible reasoning can solve problems and paradoxes occurring in monotonic deontic logic (e.g., Nute, 1997). However, previous work usually focused on ethical "oughts" rather than epistemic "oughts." We therefore hope that this paper spurs research on defeasible reasoning with epistemic "oughts."

ACKNOWLEDGMENTS

The research which led to this article was partially supported by the Fund for Scientific Research-Flanders (FWO-V). The authors also thank an anonymous reviewer for helpful suggestions.

REFERENCES

Douven, I. (2011). A role for normativism. *Behav. Brain Sci.* 34, 252–253. doi: 10.1017/S0140525X11000471

Elqayam, S., and Evans, J. St. B. T. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Hume, D. (1739–1740). *A Treatise of Human Nature*. Available online at: http://www.gutenberg.org/ebooks/4705

Moore, G. E. (1988). *Principia Ethica*. Avilable online at: http://fair-use.org/g-e-moore/principia- ethica/ (Accessed: March, 2014).

Nute, D. (1997). *Defeasible Deontic Logic*. Dordrecht; Boston; London: Kluwer Academic Publishers.

Oaksford, M., and Chater, N. (2009). Précis of bayesian rationality: the probabilistic approach to human reasoning. *Behav. Brain Sci.* 32, 69–84. doi: 10.1017/S0140525X09000284

Oaksford, M., and Chater, N. (2011). The "is-ought fallacy" fallacy. *Behav. Brain Sci.* 34, 262–263. doi: 10.1017/S0140525X11000665

Oaksford, M., and Sellen, J. (2000). Paradoxical individual differences in conditional inference. *Behav. Brain Sci.* 23, 691–692.

Pigden, C. R. (2010). *Hume on Is and Ought*. Basingstoke; Hampshire; New York: Palgrave Macmillan.

Pollock, J. L. (1987). Defeasible reasoning. *Cogn. Sci.* 11, 481–518. doi: 10.1207/s15516709cog1104_4

Pollock, J. L. (1992). How to reason defeasibly. *Artif. Intell.* 57, 1–42. doi: 10.1016/0004-3702(92)90103-5

Quintelier, K. J. P., Van Speybroeck, L., and Braeckman, J. (2011). Normative ethics does not need a foundation: it needs more science. *Acta Biotheor.* 59, 29–51. doi: 10.1007/s10441-010-9096-7

Schneider, S. L. (2000). An elitist naturalistic fallacy and the automatic-controlled continuum. *Behav. Brain Sci.* 23, 695–696. doi: 10.1017/S0140525X00553436

Schroyens, W. (2009). On is and ought: levels of analysis and the descriptive versus normative analysis of human reasoning. *Behav. Brain Sci.* 32, 101–102. doi: 10.1017/S0140525X09000478

Stanovich, K. E., and West, R. F. (2000). Advancing the rationality debate. *Behav. Brain Sci.* 23, 701–717. doi: 10.1017/S0140525X00623439

Schurz, G. (1997). *The Is-Ought Problem—An Investigation in Philosophical Logic*. Dordrecht; Boston; London: Kluwer Academic Publishers. Avilable online at: http://www.springer.com/philosophy/logic+and+philosophy+of+language/book/978-0-7923-4410-0

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 07 May 2014; published online: 27 May 2014.

*Citation: Quintelier KJP and Zijlstra L (2014) How (not) to argue about is/ought inferences in the cognitive sciences. Front. Psychol. **5**:503. doi: 10.3389/fpsyg.2014.00503*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

In defence of soft normativism

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 05 November 2014
doi: 10.3389/fpsyg.2014.01269

The intersection between Descriptivism and Meliorism in reasoning research: further proposals in support of 'soft normativism'

***Edward J. N. Stupple*[1] ** and *Linden J. Ball*[2]**

[1] *Division of Psychology, University of Derby, Derby, UK*
[2] *School of Psychology, University of Central Lancashire, Preston, UK*

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Meredith Ria Wilkinson, De Montfort University, UK
Shira Elqayam, De Montfort University, UK
Rakefet Ackerman, Technion-Israel Institute of Technology, Israel

****Correspondence:***
Edward J. N. Stupple, Centre for Psychological Research, University of Derby, Kedleston Road, Derby, UK
e-mail: e.j.n.stupple@derby.ac.uk

The rationality paradox centers on the observation that people are highly intelligent, yet show evidence of errors and biases in their thinking when measured against normative standards. Elqayam and Evans' (2011) reject normative standards in the psychological study of thinking, reasoning and deciding in favor of a 'value-free' descriptive approach to studying high-level cognition. In reviewing Elqayam and Evans' (2011) position, we defend an alternative to descriptivism in the form of 'soft normativism,' which allows for normative evaluations alongside the pursuit of descriptive research goals. We propose that normative theories have considerable value provided that researchers: (1) are alert to the philosophical quagmire of strong relativism; (2) are mindful of the biases that can arise from utilizing normative benchmarks; and (3) engage in a focused analysis of the processing approach adopted by individual reasoners. We address the controversial 'is–ought' inference in this context and appeal to a 'bridging solution' to this contested inference that is based on the concept of 'informal reflective equilibrium.' Furthermore, we draw on Elqayam and Evans' (2011) recognition of a role for normative benchmarks in research programs that are devised to enhance reasoning performance and we argue that such Meliorist research programs have a valuable reciprocal relationship with descriptivist accounts of reasoning. In sum, we believe that descriptions of reasoning processes are fundamentally enriched by evaluations of reasoning quality, and argue that if such standards are discarded altogether then our explanations and descriptions of reasoning processes are severely undermined.

Keywords: rationality paradox, normativism, radical relativism, descriptivism, soft normativism, reflective equilibrium, individual differences, reasoning

INTRODUCTION

The rationality paradox (e.g., Evans and Over, 1996) centers on the observation that people are demonstrably highly intelligent, yet simultaneously show evidence of numerous errors and biases in their thinking, reasoning and deciding when measured against normative standards associated with formal, logical systems or probability theory. This rationality paradox has emerged from a paradigm that sets descriptions of what human thinking 'is' against prescriptions of what human thinking 'ought' to be. This paradigm is based around what Elqayam and Evans (2011) describe as 'prescriptive normativism,' and can be traced back to pioneering research on systematic errors in reasoning by Wason (1966) and Tversky and Kahneman (1974). The paradigm is also central to the more recent program of individual differences research by Stanovich and West (2000) and Stanovich et al. (2010) that plays squarely into a Meliorist agenda, which views people's reasoning as being amenable to improvement through training and education. Elqayam and Evans (2011), however, have presented a powerful critique of 'normativism' in reasoning research, whether of the prescriptive variety favored by Meliorists or of the 'empirical' variety, favored by Panglossian theorists (e.g., Oaksford and Chater, 2007), who propose that human reasoning is *a priori* rational, having been forged by adaptive evolutionary forces that have patterned fitness-relevant characteristics that enable effective goal attainment.

Elqayam and Evans' (2011) critique argues that both prescriptive and empirical normativism invite researchers to make a logically contested 'is–ought' inference. In the case of prescriptive normativism, when there are competing normative accounts then empirical 'is' evidence is inevitably called upon as a basis for arbitration, giving rise to a clear case of is–ought reasoning. For example, Stanovich and West (2000) have proposed that the reasoning of the most cognitively able respondents can arbitrate between opposing normative accounts (cf. Stich and Nisbett, 1980). In the case of empirical normativism, the is–ought inference arises by virtue of the Panglossian view that 'average' or 'modal' responses that occur on reasoning tasks are an index of normative reasoning (e.g., Cohen, 1981; cf. Oaksford and Chater's, 2007, rational analysis approach). Elqayam and Evans (2011) contend that normativism in both of these guises should be strictly avoided given that the dubious is–ought inference that it invokes fosters misunderstandings and obstructs sound theorizing. They instead advocate a descriptivist analysis of reasoning as the only viable way forward for the study of high-level cognition. Elqayam and Evans (2011) further suggest that there is an acceptable role for 'formal systems' in theory development when such formal systems are

applied in a non-evaluative manner. In this respect they propose that logical systems or probability theory can be useful in providing a 'computational-level' analysis (Marr, 1982) or a 'competence' theory (Chomsky, 1965) in terms of offering structural descriptions of people's abstract knowledge that are nevertheless 'value-free.' It is also noteworthy that Elqayam and Evans (2011) propose that normative approaches *can* be useful in one very restricted sense, that is, when the researcher has an applied objective "to improve thinking (rather than understand it)" (p. 242), since it is then necessary to have criteria that can distinguish good thinking from bad thinking.

In this paper we set out to defend an approach that can be seen as a middle ground between a descriptivist perspective and a stance that is based on prescriptive normativism. The approach that we advocate has been dubbed 'soft normativism' by Evans and Elqayam (2011), and is a position that sees a role for normative evaluation in reasoning research alongside the pursuit of descriptive research goals. Our argument (cf. Stupple and Ball, 2011) proposes that normative theories have considerable value for formulating and testing hypotheses, provided that researchers: (1) remain alert to the philosophical quagmire of more radical forms of relativism (e.g., Stich, 1990; Elqayam, 2012); (2) are mindful of the biases and pitfalls that can arise from drawing on normative accounts of reasoning (see Elqayam and Evans, 2011); and (3) ensure that they engage in a focused analysis of the processing approach adopted by *individual* reasoners when confronted with reasoning tasks (cf. Stanovich and West, 2000).

In this paper we will examine the position of soft normativism in the context of dual-process theories of reasoning, individual differences in reasoning and attempts to ameliorate reasoning 'defects.' Our conclusion is that descriptions of reasoning processes are fundamentally enriched by evaluations about the quality of that reasoning. As such, soft normativism is, we suggest, a reasonable pragmatic position to take when both judging reasoning and when formulating theoretical accounts. In developing our argument we also address how soft normativism can circumvent the contested is–ought inference by means of a well-recognized bridging solution (see Evans and Elqayam, 2011) that is based on the concept of reflective equilibrium (Goodman, 1965). We extend the notion of reflective equilibrium to capture the way in which the reasoning behavior of naïve individuals changes when they are provided with extensive opportunities to practice their reasoning.

We additionally propose that the distinction between the applied science of 'improving thinking' versus the pure science of 'understanding thinking' it not a clear-cut dichotomy of the kind that Evans and Elqayam might like to envisage. Because of this overlap our own position sits at the intersection between Meliorist and descriptivist research agendas. On the one hand we believe that to inform efforts to enhance reasoning and argumentation we must have a good understanding of underlying reasoning processes, since this will aid our explanation of why some individuals are better at making arguments or drawing inferences than others. On the other hand, the converse relationship is also important, since understanding the way in which Meliorist approaches are effective in enhancing reasoning can supplement our theoretical understanding of underlying reasoning processes. In other words, studying the improvements that can arise in thinking, reasoning or judgment through training or educational interventions allows researchers to draw important comparisons between what people can achieve as a result of such external guidance (coupled with their own reflective process) and what people can achieve through a spontaneous process. This contrast between a 'sophisticated' versus 'naïve' reasoning process is psychologically interesting and arises directly from Meliorist researchers' attempts to align reasoning with external, normative benchmarks.

STRONG NORMATIVISM AND RADICAL RELATIVISM

Extreme views at either end of the normativism–descriptivism spectrum are beset with problems. A strong Panglossian normativist such as Cohen (1981) must be able to demonstrate that all errors of reasoning can be explained away through lapses of attention, misunderstandings of instructions or (ecologically invalid) cognitive illusions. Stanovich and West (2000), however, have convincingly demonstrated that if errors are predominantly the result of lapses of attention then such errors within tasks should be uncorrelated since these lapses will be randomly distributed, and likewise performance across tasks should also be uncorrelated for the same reason. A wealth of evidence, however, has shown this not to be the case, with systematic errors and biases being demonstrated within tasks and with clear correlations arising between various reasoning and judgment tasks. The correlations in performance across tasks are not without exceptions[1], but do nevertheless provide evidence that is problematic for the view that the so-called 'normative–descriptive gap' (Stanovich and West, 2000) can be explained within the framework of pure normativism.

Moreover, Stich (1990) presents further challenges for adherents of strong normative positions with his concept of 'cognitive pluralism': that there is more than one good way to reason. Stich illustrates this view with comparisons to alternative cultures that may not share Western ideals about particular normative systems, and he further extends this position (via Goldman, 1986) to ask whether a normative system must hold in all possible worlds (or at least 'normal worlds') if it is truly universal. If we concur with Stich's argument then we have stepped onto the slippery slope to radical relativism and have accepted that there is no universal benchmark to judge inferences (or to make judgments about judgments) that can apply to all contexts. If we continue all the way to the bottom of this slippery slope then we reach the anarchic conclusion that all inferences and choices are equal. Buckwalter and Stich (2011) note, however, that the concept of cognitive pluralism made little headway initially, and they suggest that this was because there was no compelling evidence that it is 'psychologically' possible for people to have significantly different reasoning competences. They concur, however, that Stanovich's (1999) research program on individual differences in reasoning goes a fair way toward demonstrating that there are indeed a range of such competences. Although this does not indicate that all possible inferences are justifiable, it nevertheless indicates that a

[1] It might be argued that those tasks that do not correlate with other tasks are examples of ones that give rise to 'cognitive illusions,' as described by Cohen (1981), or else are tasks where specific 'mindware' (i.e., specialized cognitive rules or strategies; see Stanovich, 2009) is more important than more general reasoning ability.

degree of relativism may be undeniable when it comes to describing reasoning competence. The concept of cognitive pluralism advanced by Stich not only poses problems for using normative benchmarks as standards to judge thinking and reasoning, but also has the potential to be challenging for computational-level or competence-based descriptive accounts of thinking and reasoning by virtue of the need for an explanation of why such varied competences arise.

We nonetheless defend a moderate relativism by noting that the 'slippery slope' argument is a well-known fallacy such that we can make progress by recognizing the limitations of accounts that assume normative benchmarks and cognitive universality, while also acknowledging some of the important issues that normative views raise in defense of human rationality. Indeed, Samuels and Stich (2004) have likewise argued for a 'middle way' when conceiving of human rationality in the context of dual-process theory, which they see as offering an escape from the rationality paradox. It is in a similar spirit that we argue for the application of a softer normativism when engaging in reasoning research. The soft normativism that we advocate is admittedly a few steps down the slippery slope toward relativism in that it recognizes a role for context and participant knowledge in judging the efficacy or appropriateness of an inference. Nevertheless, our proposed soft normativism still places considerable value on people's ability to produce valid inferences in response to reasoning and decision making tasks.

The crux of our position is that while we endorse the use of normative standards as the basis for a descriptively oriented computational level of analysis (or what might also be viewed as a 'competence theory' of reasoning; Elqayam and Evans, 2011), and while we also acknowledge that from a descriptivist perspective there is no additional value in regarding deviations from these standards as 'errors,' we still believe that normative standards can benefit the applied study of reasoning provided that they are deployed sensibly. We advocate the use of normative standards – in accordance with Elqayam and Evans (2011) – where the goal of research is to enhance reasoning, argumentation or judgment so as to align it with external benchmarks of quality. Certainly a primary goal of Meliorist researchers is to increase the proportion of people who avoid bias and endorse some set of external normative standards (i.e., is the desire is to ensure that people reason as well as they are able to). This Meliorist agenda represents a substantial research program within the discipline of Cognitive Science that is either followed explicitly (e.g., Stanovich, 2011) or else implicitly (e.g., Ball, 2013b). We contend that it is simply not possible to have such a Meliorist agenda without some notion of what constitutes 'good' thinking. We further assert that while there can be a range of normative theories that apply to a particular reasoning or decision making domain (i.e., we accept that these standards can be controversial), they still offer us a guide as to what constitutes good reasoning or judgment. Thus, when a Meliorist researcher succeeds in enhancing thinking, this change in performance (or even the capacity to change) needs to be compatible with a computational-level explanation (i.e., the descriptivist researcher needs to be able to explain the Meliorist researcher's findings). In this latter respect we believe that a soft normativist compromise is what is required to allow for a mutually beneficial symbiosis between Meliorist and descriptivist agendas.

PITFALLS WHEN DRAWING ON NORMATIVE THEORIES

Theorists such as Cohen (1981) have argued that the reasoning research paradigm has not been particularly charitable to participants over the years, with a tendency to present 'trick' questions with minimalist instructions to naïve individuals. Judging non-normative responses as indicative of irrationality on this basis is, he proposes, difficult to justify. Evans (2007) has argued that researchers such as Cohen who attempt to defend human rationality have tended to do so by appealing to three key problems with the attribution of irrationality to reasoners: (1) the normative system problem; (2) the interpretation problem; and (3) the external validity problem (see also Evans, 1993; Evans and Over, 1996). The normative system problem is that researchers are simply applying the wrong normative standards when judging participants' task performance, such that if the correct normative system were applied then behavior could be re-classified as rational. It is worth noting that there are many logical systems (e.g., see Garson, 2014) and that this diversity has provoked debate about which normative system is the 'correct' one in any particular reasoning context. Such diversity can also prompt interesting questions as to the characteristics of individuals who endorse differing benchmarks when multiple standards are available. For the Meliorist, however, it is inevitable that there will be a degree of 'satisficing' when selecting a normative standard against which to judge reasoning or decision making, since such standards can be debatable and can develop and change through cultural evolution as tools of rationality. As Stanovich (2011) notes: "... there is no idealized human 'rational competence' that has remained fixed throughout history" (p. 269).

The interpretation problem explains participants' deviations from normative theory not in terms of the application of faulty reasoning processes but instead in terms of participants adopting alternative mental representations of problem information to that intended by researchers. The external validity problem, which is closely allied with Cohen's (1981) argument noted above, is that the tasks that researchers select in order to demonstrate human irrationality are not at all representative of the tasks that arise in real-world contexts, which tend to be associated with normatively accurate reasoning. Evans (2007) argues that the interpretation problem and the external validity do not hold up to close scrutiny because they fail to offer 'complete' accounts of the discrepancy between normative benchmarks and actual behavior. We would counter, however, that neither approach needs to offer a comprehensive account of normative–descriptive discrepancies so long as each approach can offer up explanations of at least some of the relevant data. Take, for example, the interpretation problem as discussed by Evans (2007). There is a good degree of consensus in the literature (e.g., Stanovich and West, 2000) that there are individual differences in cognitive ability and thinking dispositions that influence reasoning. There is, moreover, evidence of individual differences in the *interpretation* of elements of the reasoning scenarios and vignettes that participants tackle in the laboratory (e.g., Roberts et al., 2001; Stenning and Cox, 2006). For example, if an individual fails to interpret the quantified

assertion *Some A are B* as possibly meaning *All A are B*, then this invites the assumption that *Some A are not B* is also true (Newstead and Griggs, 1983)[2]. Newstead (1989) demonstrated that quantifier interpretation can indeed influence performance in some circumstances, and Roberts et al. (2001) showed that while there can be 'errors' based on interpretation, these vary according to the complexity of the task. Stenning and Cox (2006) further revealed that individual differences in the interpretation of quantifiers result in differing patterns of responses. In sum, it seems important to acknowledge that the interpretation problem is a very real one, even if it does not provide a complete explanation of deviations from normative benchmarks in all situations and even if explaining the findings that arise in studies is not always straightforward.

This interpretation problem in reasoning research also has implications for explaining reasoning accuracy in the context of dual-process theories that invoke a distinction between rapid, effortless and intuitive 'Type$_1$' processes and slow, effortful and analytic 'Type$_2$' processes (e.g., Evans and Stanovich, 2013). The predominance of naive participants in reasoning studies means that some task misinterpretation is inevitable, which confounds any inferences that researchers might want to make either about non-normative responding reflecting Type$_1$ processing or about normative responding reflecting Type$_2$ processing (see also Thompson, 2011). In the latter case, for example, if quantifiers in syllogistic reasoning tasks are misinterpreted then non-normative responding might still be based on effortful Type$_2$ thinking. In this respect we are reminded of Smedslund's (1990) critique of Kahneman and Tversky's (1972) heuristics and biases paradigm, whereby he argued that we cannot decide if someone has reasoned logically unless we assume they represented the premises as the experimenter intended, and likewise we cannot judge whether someone represented the premises as intended unless we assume they reasoned logically. This circularity continues to be an issue when equating normative responses with Type$_2$ processing (e.g., see Evans, 2012). For example, someone could employ a normative goal (i.e., to reason logically) and pursue this goal with great effort using Type$_2$ processing, and yet still offer a non-normative response because they are unaware of the need for a 'non-pragmatic' interpretation of a quantifier (i.e., an interpretation that is inconsistent with everyday usage). In fact, Noveck and Reboul (2008; see also Bott and Noveck, 2004) have shown that effortful processing is required to narrow *Some* to *Some but not all*, which means that in some cases a pragmatic interpretation may require more Type$_2$ processing than a normative response.

Whilst these aforementioned issues might be seen to undermine entirely any agenda that attempts to align participants' responses with normative theories, we would argue instead that such issues simply alert researchers to the need for more cautious interpretation of reasoning data. Indeed, we would go a step further and propose that such issues can guide the careful design of experiments in the first place so that they can accommodate the way in which participants are likely to engage in pragmatic interpretations of information. An example of just such an approach comes from a study by Schmidt and Thompson (2008), who used the quantifier *At least one and possibly all* instead of *Some* within given premises and found that participants were facilitated in giving normative responses. Whilst the deductive paradigm and instantiations of it, such as the belief-bias paradigm[3] (e.g., Evans et al., 1983; Stupple and Ball, 2008), continue to be important test-beds for dual-process accounts of reasoning, we advocate the increased utilization of pragmatically interpretable quantifiers (or else instructions regarding how quantifiers should be interpreted) in order to increase precision when deciphering apparent variations between normative benchmarks, descriptions of performance and the alignment of outputs with Type$_2$ processing.

THE IMPORTANCE OF DATA TRIANGULATION IN EVALUATING THE NORMATIVE BASIS OF REASONING

Given that pragmatic interpretations and responses can explain some (but not all) deviations from normative standards, we believe that it is increasingly important to include the triangulation of measures (e.g., response types, processing times, and confidence judgments) in any empirical studies that are examining the nature of reasoning, including its normative basis and possible dual-process components. In this respect it has been encouraging to see a burgeoning over the past decade or so in the use of 'multi-method' approaches in reasoning research (for good examples of such multi-method studies see Quayle and Ball, 2000; Thompson et al., 2003, 2011a,b, 2013; De Neys, 2006; Stupple and Ball, 2007, 2008; De Neys and Glumicic, 2008; Prowse Turner and Thompson, 2009; De Neys et al., 2011; Stupple et al., 2011). Particularly valuable insights into the nature and time-course of reasoning processes can be gained by examining think-aloud protocols that are acquired from participants who are tackling reasoning problems (e.g., Evans et al., 1983; Lucas and Ball, 2005; Wilkinson et al., 2010) as well as by analyzing neuroimaging data collected concurrent to reasoning performance (e.g., Goel and Dolan, 2003; Luo et al., 2013)[4]. Houdé (2007) has, in fact, recently argued that "... one of the crucial challenges for the cognitive and educational neuroscience of today is to discover the brain mechanisms that enable shifting from reasoning errors to logical thinking" (p. 82). The challenge that Houdé refers to clearly requires a major drive toward the increasing deployment of triangulating measures that attempt to understand the neural underpinnings associated with the transition that people are able to make toward normative responding through training and education. A recent

[2]This is an example of an issue of scalar implicature, as discussed by Grice (1975), whereby there is a clash between the quality of the information provided and then quantity of information provided. In a cooperative social exchange the use of the quantifier 'Some' when it is possible to use the quantifier 'All' violates Gricean maxims of effective communication.

[3]Belief-bias is a pervasive tendency in reasoning to accept believable conclusions more frequently than conclusions that contradict beliefs, irrespective of the logical validity of conclusions (see Evans et al., 1983, for pioneering research on this phenomenon that also established the standard 'belief-bias paradigm' that inspired most subsequent research).

[4]Another interesting methodology that is being used increasingly in the study of reasoning concerns the measurement and analysis of autonomic arousal (e.g., De Neys et al., 2010; Morsanyi and Handley, 2012), which appears to reveal participants' implicit awareness of reasoning conflicts (e.g., between the logical status and belief status of conclusions).

example of such an approach comes from Luo et al. (2014) who demonstrated differences in activation for the left inferior frontal gyrus, left middle frontal gyrus, cerebellum, and precuneus for a group of highly belief-biased participants who had subsequently received logic training and switched to logic-based responding.

A further monitoring approach for examining the dynamic aspects of reasoning that we are particularly enthusiastic about is to deploy eye-tracking (e.g., Ball et al., 2003, 2006) to determine the moment-by-moment attentional shifts in processing that arise when participants attempt the visually presented problems that are typically used in reasoning studies (see Ball, 2013a, for a recent summary of key findings deriving from eye-tracking research in reasoning). Eye-tracking studies have, we contend, provided some of the most compelling evidence to date that $Type_2$ analytic reasoning that is attuned to normative principles plays an important role in determining whether heuristically cued cards are subsequently selected or rejected in the Wason four-card selection task (see Evans and Ball, 2010). Likewise, eye-tracking studies of belief-bias effects (e.g., Ball et al., 2006) have been influential in revealing that people spend longer reasoning about 'conflict' syllogisms, where conclusion validity and believability are in competition (i.e., those with invalid-believable conclusions and valid-unbelievable conclusions), relative to 'non-conflict' syllogisms, where conclusion validity and believability concur (i.e., those with valid-believable and invalid-unbelievable conclusions). The evidence that conflict problems take longer to process than non-conflict problems is viewed by Stupple and Ball (2008) as indicating that participants are 'sensitive' to the fact that the logic of a conclusion and its belief status are in opposition such that extra processing effort has to be allocated to resolving the conflict. Such findings resonate with recent proposals that have been forwarded by De Neys (2012), who suggests that people's indirect sensitivity to the normative status of presented conflict conclusions is indicative of their possession of an 'intuitive logic' (a $Type_1$ process) that functions implicitly and in parallel to implicit heuristics (also $Type_1$ processes) to signal the need for $Type_2$ processing. De Neys (2014) presents some clarifications about the role of these controversial 'gut-feelings' in shaping the way participants respond to conflict problems and asserts that whether or not we endorse his 'logical intuition' proposal we can certainly question the idea that $Type_1$ responses can typically be attributed to a failure in conflict detection. We are mindful, however, of the calls from Singmann et al. (2014) for the application of the most rigorous, scientific approach possible when examining such 'extraordinary' claims as the existence of an intuitive logic (see also Klauer and Singmann, 2013).

One important area of eye-tracking research in the reasoning domain that is currently gaining increased attention concerns the analysis of eye-movement metrics that are directly linked to people's *comprehension* of visually presented logical statements. For example Stewart et al. (2013) deployed eye-tracking to examine how readers process "if ... then" statements used to communicate conditional speech acts such as promises (which require the speaker to have perceived control over the consequent event) and tips (which do not require perceived control). Various eye-tracking measures showed that conditional promises that violated expectations regarding the presence of speaker control resulted in processing disruption, whereas conditional tips were processed equally easily regardless of whether speaker control was present or absent. Stewart et al. (2013) concluded that readers make very rapid use of pragmatic information related to perceived control in order to represent conditional speech acts as they are read. These kinds of on-line studies of 'reasoning as we read' (see also Haigh et al., 2013) seem likely to open up many new possibilities for advancing an understanding of reasoning processes by providing converging empirical evidence to help arbitrate between competing theoretical accounts.

Overall, we contend that without alternative, convergent measures of reasoning that extend well beyond mere response choices we have no direct gage of the nature and time-course of reasoning, such as whether the cognitive processing that participants deploy is slow and effortful or fast and intuitive. Furthermore, simply knowing that responses are consistent with normative benchmarks is clearly insufficient to claim that $Type_2$ thinking is involved (e.g., Evans and Stanovich, 2013; see also Evans, 2012, for important arguments and evidence in this respect). A recent illustration of this point comes from Stupple et al. (2011), who demonstrated correlations between response times and normative responding in a belief-bias paradigm, with increased response times to invalid-believable problems being indicative of increased normatively aligned performance. Thus, those participants who exhibited longer response times where there was a conflict between belief and logic, and who identified invalid-believable conclusions as 'possibly true' rather than 'necessarily true' (which requires a more complex understanding of *Some ... are not ...* than the standard pragmatic interpretation), appeared to possess the requisite cognitive resources and motivation to search for counterexamples. This meant that these participants were more likely to respond normatively to belief-oriented problems in general, and not just to the invalid-believable conflict items.

Similarly, Stupple et al. (2013) investigated 'matching bias' in syllogistic reasoning from a dual-process perspective. Matching bias is the phenomena whereby responses are simply matched to terms mentioned in a rule or are based on the surface features of premises, in either case being based on a 'non-logical' process (e.g., see Evans and Lynch, 1973; Wetherick and Gilhooly, 1995). In Stupple et al.'s (2013) study the surface features of problems were manipulated so as to be either congruent with or orthogonal to the logic of the presented conclusions. Performance was then judged based on whether it aligned with the surface features of the problems or with normative responses as determined by formal logic. This experimental set-up is much like the belief-bias paradigm, where conclusion believability and validity either concur or conflict. To manipulate the surface features of problems Stupple et al. (2013) used premises and conclusions that were matched or mismatched in terms of the presence of double negated quantifiers (e.g., *No A are not C*) or in terms of the presence of standard affirmative quantifiers (e.g., *All A are C*). Using this paradigm Stupple et al. (2013) revealed some important parallels between their results and findings deriving from studies of belief-bias. One key parallel concerned the observation that 'conflict' problems in both paradigms show inflated response times relative to non-conflict

problems (cf. Thompson et al., 2003; Ball et al., 2006; Stupple and Ball, 2008; Stupple et al., 2011), which is entirely in line with dual-process predictions and attests to the value of obtaining response-time data as a way to inform theorizing. Stupple et al.'s (2013) study also revealed that the supposedly 'intuitively obvious' deduction of *double negation elimination* (see Rips, 1994, pp. 112–113) was demonstrably unintuitive for a number of participants, who showed increased response times to problems involving such negations.

Perhaps of more pertinence to the present discussion are Stupple et al.'s (2013) findings from the same study that contrasted with what has previously been observed for problems within the standard belief-bias paradigm, particularly in relation to correlations between response times and normative response rates. In particular, valid non-matching 'conflict' problems actually revealed an association between normative responding and *faster* responses, which is distinct from what is seen in belief-bias research, where valid-unbelievable conflict items show an association between normative responding and *slower* responses (e.g., Stupple et al., 2011). To explain this discrepancy Stupple et al. (2013) proposed that motivated participants who do not possess the double elimination rule (or who have difficulty applying it) might engage in a misdirected and slow analytic process to find a matching-consistent answer (see Stupple and Waterhouse, 2009; Stupple et al., 2013), whereas participants who eliminate the double negation are confronted with little cognitive demand in identifying that the conclusion is necessarily true such that they can rapidly respond normatively. We suggest that without the reference point that normative benchmarks offer, such idiosyncrasies in individual responding may well pass unnoticed. The combination of cognitive effort, quantifier interpretation and cognitive disposition demonstrate the increasing importance of individual differences approaches in reasoning research and also illustrate the utility of having normative benchmarks as a point of comparison.

In the next section we discuss in more detail the value of adopting an individual differences perspective on reasoning strategies whilst also further examining the way in which normative reference points can benefit an understanding of reasoning data. First, however, we take a brief detour into another area of contemporary reasoning research that also exemplifies the importance of methodological triangulation, that is, research on metacognition and reasoning – or so-called 'meta-reasoning' (for a recent review see Ackerman and Thompson, 2014; for pioneering conceptual work see Thompson, 2009). This growing research topic is concerned with the processes that 'regulate' reasoning, for example, by setting goals, deciding among strategies, monitoring progress and terminating processing. The meta-reasoning framework is predicated on the assumption that people are generally motivated to attempt to provide 'right' answers to reasoning problems. Indeed, meta-reasoning is centrally concerned with an individual engaging in processes such as determining how much effort to apply to the problem, assessing whether a solution that they have generated is correct, and deciding whether to initiate further processing if a putative solution seems in some way inadequate (Thompson, 2009; Ackerman and Thompson, 2014). As a case in point, Ackerman and Thompson (2014) suggest that the very first decision that that a reasoner should make is that of whether to attempt a solution at all, since the individual might determine that the amount of effort they need to apply to achieve a solution is greater than the perceived benefit of solving the problem (cf. Kruglanski et al., 2012). Ackerman and Thompson (2014) suggest that such 'Judgments of Solvability' are likely to be based on a range of factors, including beliefs about the task at hand, prior experience of solving similar problems, as well as surface-level cues within the problem itself that might signal difficulty, such as the ease with which the problem can be mentally represented (e.g., Quayle and Ball, 2000; Stupple et al., 2013) or the perceived coherence amongst problem elements (e.g., Topolinski and Reber, 2010; Topolinski, 2014).

As can be seen in relation to Judgments of Solvability, the meta-reasoning framework presupposes that people do not have direct access to their underlying reasoning processes, but instead base their monitoring and regulation judgments on their experience with similar problems as well as on available cues associated with the problem being tackled. One particularly important cue is that of 'fluency,' which is the ease or speed with which a solution to a reasoning problem comes to mind (e.g., Alter and Oppenheimer, 2009; Ackerman and Zalmanov, 2012). Thus, an individual will generally view an initial response that is produced fluently as being accurate, whereas an initial response that is difficult to generate will give rise to a sense of unease in relation to its accuracy, often triggering further processing effort. Importantly, such heuristic cues to accuracy may not be valid predictors of normative correctness, leading to some striking dissociations between participants' response confidence and normative standards of accuracy (e.g., see Shynkaruk and Thompson, 2006; Prowse Turner and Thompson, 2009; De Neys et al., 2013). Thompson (2009) and Thompson et al. (2011b, 2013) have gone beyond the basic concept of answer fluency in their theorizing to suggest that such fluency mediates a judgment that they term 'Feeling of Rightness.' It is this Feeling of Rightness judgment that then acts as a metacognitive trigger, either: (1) terminating processing in cases where a $Type_1$ process has readily produced a rapid, intuitive answer that is attributed to be correct; or (2) switching from $Type_1$ to $Type_2$ processing in cases where the initial, intuitive answer is associated with a low Feeling of Rightness and is therefore attributed to be potentially incorrect (see Ackerman, 2014, for further evidence and model development regarding people's time investment in reasoning).

In sum, recent evidence gives clear grounds for viewing meta-reasoning judgments as playing a crucial role in monitoring and regulating on-going reasoning, such that intermediate confidence or 'rightness' assessments determine the amount of subsequent effort that reasoners invest in a task (Ackerman, 2014). The methodology underpinning this meta-reasoning research is based on a rich triangulation of measures, including various forms of confidence judgments as well as processing times and normative response accuracy. Analyzing confidence ratings in conjunction with other measures also seems advantageous in terms of distinguishing between normatively incorrect answers that participants 'expect' to be correct with high probability and wild guesses (i.e., responses made with particularly low confidence), which perhaps

reflect task abandonment and may therefore be of less theoretical interest.

We suggest that the evident tendency in meta-reasoning research to evaluate participants' responses against normative benchmarks such as logic, suggests that this emerging research tradition has a strong normative orientation, which is also bolstered by the inherent assumption underpinning the approach that participants are generally striving to produce 'right' answers to problems. Notwithstanding our view that normative considerations have an important role to play in emerging research on meta-reasoning, we do nevertheless concur with Thompson's (2011) argument that simply knowing that a final outcome is normative tells us virtually nothing about underlying mechanisms. At the same time, however, we believe that a combination of process-oriented analyses together with the normative assessment of outcomes provides for a maximally rich and meaningful approach to reasoning research, especially when combined with studies of the roles of learning, practice and feedback in reasoning, as discussed below. These themes tap directly into a Meliorist research agenda, where evaluations of normative correctness are crucial. In this respect we look forward to further research using measures such as Judgment of Solvability and Feeling of Rightness in the context of training reasoning through instructions, practice, and feedback. We believe that such work could inspire new insights into the monitoring and regulatory processes that lead to both normative and non-normative reasoning responses, whilst also benefiting applied research on improving reasoning (see Ackerman and Thompson, 2014, for discussion of numerous real-world domains that could be enhanced through such research, such as innovative product design and financial decision making).

THE IMPORTANCE OF FOCUSING ON INDIVIDUAL DIFFERENCES

Since the seminal research of Gilhooly et al. (1993), Roberts (1993), and Ford (1995) there has been a snowballing of individual differences studies in reasoning research, perhaps best exemplified by the work of Stanovich and colleagues (e.g., Stanovich and West, 2000). The question of why some participants respond in accordance with normative standards more frequently than others forms an important research agenda in its own right, but the ability to account for individual differences within a particular theoretical framework is increasingly part of the debate in a range of reasoning research paradigms (e.g., Stupple et al., 2011; Trippas et al., 2013). Nickerson (2008) argues that determining which normative system is the best one in a given context is often an uninteresting issue, unless it also happens that aligning cognitive processing with the normative system in question also correlates with something that people care about. A strong supporter of a normativist agenda could argue that since Stanovich and colleagues have demonstrated correlations between tasks from the reasoning and decision making literature with things that are prized – such as SAT scores – then it is possible to believe that there is something valuable in adhering to these normative standards. If our instrumental goals are to gain a place at a prestigious university or to score well on an employer's recruitment test of cognitive ability, then reasoning and deciding in accordance with normative benchmarks can be an instrumental goal, at least for some participants some of the time[5].

There is, nevertheless, much debate concerning the issue of how normative standards can be derived in the first place. The concept of 'reflective equilibrium' is central to this debate, and was a notion that was advanced by Goodman (1965), who argued that as the rules of deduction are determined by accepted deductive practice then good deductive rules are retained and poor deductive rules that lead to poor inferences are dropped. This is a rigorous circular process that is engaged in by philosophers and logicians in developing normative standards of inference. This concept was further developed by Cohen (1981) in the context of the rationality paradox. The idea is that normative theory and descriptive evidence can justify each other by being brought into coherence such that there is an alignment between norms and behaviors. As Elqayam and Evans (2011) note, reflective equilibrium is a 'bridging solution' to the notorious is–ought problem since it presupposes that full coherence is entirely possible between norms and behavior inasmuch as they become mutually justificatory.

Of course, the proposal that reflective equilibrium can offer a route to deriving appropriate normative benchmarks is not without its critics, with Stich (1990), for example, emphasizing that it has the potential once again to lead down the slippery slope to radical relativism. Stich argues that the gambler's fallacy and base rate neglect pass many people's tests of reflective equilibrium, which indicates that the principle can be flawed as a means of justifying inferences. Stich also demonstrates that the issue cannot be solved if we impose restrictions on the people whose reflective equilibrium is considered to be sufficiently rigorous to serve as a justification, since even experts could "end up endorsing a nutty set of rules" (p. 86). There may also be cultural and interpersonal differences in assessing the justification of an inference that yield different benchmarks in different contexts. For many, Stich's critique would appear terminal for the use of reflective equilibrium as a means of justifying universal norms for inference. Nevertheless, his critique does not entirely rule out the application of similar principles by individuals in justifying their own inferences and judgments. Indeed, it is possible that participants can engage in an informal process analogous to reflective equilibrium in establishing how they should respond to reasoning tasks.

Recent findings by Ball (2013b) advance this aforementioned concept of 'informal reflective equilibrium' by indicating that participants will, through repeated reasoning practice, develop their own benchmarks for accuracy. This observation seems further to support a moderate relativism that functions hand-in-glove

[5]The concept of 'instrumental goals' relates to Evans and Over's (1996) notion of Rationality$_1$, that is, 'instrumental' or 'pragmatic' rationality, defined in terms of thinking or deciding in a way that is generally reliable and efficient for achieving one's personal goals. As such, Rationality$_1$ extends to genetically hard-wired procedures and experientially acquired processes that are automatic and implicit in nature. Evans and Over (1996) contrast the concept of Rationality$_1$ with Rationality$_2$, with this latter type of rationality being defined in terms of acting when one has a reason for what one does that is sanctioned by a normative theory. This means that the individual is not merely complying with normative rules in an implicit manner, but is following such rules explicitly.

with soft normativism. Ball (2013b), for example, demonstrated that participants who repeatedly engaged in reasoning with belief-oriented syllogisms that are known to be susceptible to a non-logical belief-bias became steadily more normatively justified in their responding over time. This trend toward increased normative responding was seen to arise even more quickly amongst those receiving feedback regarding the logical appropriateness of their decisions. Ball's findings suggests that through mere engagement and increasing familiarity with reasoning tasks people can self-determine a strategy that can affect a logical solution. Such evidence suggests a novel perspective on reflective equilibrium that is not based so much on what the most cognitively able do or what the majority do, but which instead is based on what individuals do when provided with opportunities for practice. This type of informal or 'naïve' reflective equilibrium admittedly lacks the rigor of the approach advanced by Goodman (1965), but it nevertheless indicates that untrained participants can align themselves with normative benchmarks without explicitly knowing that they are doing so or receiving feedback indicating that this is the case. Not all participants succeed in such normative alignment, and it could be argued that there is an element of 'satisficing' entailed in this process (e.g., see Evans, 2006, 2007), whereby individual differences in cognitive ability, disposition and motivation may all play an important role.

The present claims regarding the concept of informal reflective equilibrium – as well as Ball's (2013b) empirical evidence – seem to chime with the radical idea mentioned earlier that people may have 'logical intuitions,' as demonstrated, for example, by their decreased confidence when rejecting normative responses and endorsing non-normative responses (e.g., De Neys, 2012, 2014; De Neys and Bonnefon, 2013; see also Stupple et al., 2013). In Ball's (2013b) study the steadily increasing normative responding that was observed over time by the participants who did not receive feedback might well have been shaped by a repeated sense of metacognitive dissatisfaction with proffered answers – arising from 'logical intuitions' – in cases where such answers contradicted normative benchmarks. An alternative view is that through repeated exposure to belief-biased problems, the $Type_2$ analytic process becomes better attuned to the problem structure and participants become increasingly aware of the role of counterexample models in invalidating presented conclusions, irrespective of their belief status. Such issues warrant further investigation, but a purely descriptivist approach to reasoning research would rule out the use of logic as a normative reference point when scrutinizing participants' responses and would, moreover, seem to render these avenues of investigation out of bounds, irrespective of their scientific merit.

If we disallow normative theories from being utilized to inform the development of research paradigms we believe that we are, in fact, introducing a new benchmark for conducting reasoning research that is potentially obstructive to progress. For example, if the use of counterexamples is useful for good argumentation (e.g., Weston, 2009) then it is not only important to encourage our students to consider counterexamples to improve their arguments, but also for us as cognitive psychologists to understand the processes whereby individuals become attuned to the need to consider counterexamples in order to reason better. More generally, by understanding the underlying cognitive processes, we can better inform methods for improving thinking, but this would be hampered if we were not able to make value judgments about the way that participants approach their task. As another example we again refer to the belief-bias study by Stupple et al. (2011) that we outlined previously, which demonstrated that the most normatively consistent reasoners with belief-oriented syllogisms were those who had inflated response times for a particular item type that required the consideration of counterintuitive counterexamples. Stupple et al.'s (2011) evidence reconciled the descriptivist 'selective processing theory' of Evans (2000) with a previously conflicting data-set arising from a study by Stupple and Ball (2008). In addition, Stupple et al.'s (2011) evidence was informative from a Meliorist perspective, since it highlighted elements of reasoning tasks that are particularly demanding whilst also revealing individual differences in processing that correlate with solution success.

When participants engage in reasoning experiments they are likely to assume there are 'right' answers to the tasks (see the discussion above on meta-reasoning), and without giving them explicit guidance about normative standards we leave them to attain their own reflective equilibrium. Experimenters generally instruct participants what they *should* do when engaging in the task. For example, Cherubini et al. (1998) instructed participants by noting that: "Conclusions should follow from the statements only, and should be certain direct consequences of them ... You should therefore try to ignore any knowledge of what the premises are about and try to reason as if they were true" (p. 186). If experimenters direct participants to engage with a task in particular ways then there is often an explicit 'ought' as to the answers they are asked to provide. Moreover, it is generally indicated that there are correct solutions, as can be seen in the following instructions from a study by Morley et al. (2004): "This experiment is designed to find out how people *solve* logical problems ... Please take your time and be sure that you have the *logically correct answer* before deciding" (p. 8, italics added for emphasis). Even the most recent reasoning papers continue to use phrases such as, "If you judge that the conclusion necessarily follows from the premises, you should answer 'Valid,' otherwise you should answer 'Invalid' ..." (Trippas et al., 2014, p. 11). We suggest that without instructing participants that there is a correct or valid answer to a given problem it is unlikely that standard effects from the reasoning literature would arise. More generally, we contend that it is actually very difficult to envision a way in which to present 'value-free' instructions in any meaningful sense.

When presented with reasoning instructions – whether these involve explicit directives or implicit hints that there is a 'correct' solution – participants may not generate answers that conform to intended normative benchmarks, but instead may provide personally justifiable responses, based upon their understandings of the task. Some participants take longer than others over the given tasks, suggesting they have set more stringent personal thresholds of reasoning adequacy. Others may find that their intuitive responses are satisfactory (e.g., to dismiss what is unbelievable or to endorse what is intuitive). Indeed, for some there appears

to be little reasoning analysis taking place at all, as arises with the fastest responders who often seem to lack any disposition to engage in reflective, analytic, $Type_2$ thinking when confronted with reasoning tasks. In examining such individual variation we again argue that the best way to inform and enrich theoretical proposals is by triangulating a multiplicity of measures (e.g., response times, confidence judgments, and thinking dispositions) in a way that is informed by normative benchmarks (see above; cf. Ball, 2013a). Such an approach can be highly informative, provided researchers are cautious regarding disputes over such benchmarks and the dangers of directly equating normative responses with the deployment of analytic processes. The question of arbitrating between competing benchmarks is also considered by Crupi and Girotto (2014), who argue that this lies in the realm of philosophy rather than psychology and that the issue of arbitrating between competing normative standards has not played a particularly significant role in the reasoning literature. We have some sympathy with this observation, but would add that it nevertheless remains interesting and important to investigate the psychological basis for why different reasoners align with different normative standards, as in the case of Wason's (1966) selection task, where some participants appear to reason according to Oaksford and Chater's (1994) 'information gain' benchmark whilst others appear to reason according to the benchmark of propositional calculus.

On the individual differences theme we also note that since the most academically gifted tend to be those who are more cognitively able, more motivated to find the 'right' answer, and less inconvenienced by the need to engage effortful, reflective processing, then it is likely that their responses will correspond with those predicted by normative theories. This is particularly likely to be the case when those responses require additional cognitive effort and motivation, such as occasions where $Type_1$ and $Type_2$ processes come into conflict and the reasoner concords with a $Type_2$ response. The fact that the answers of these reasoners correspond with those of gifted professors of logic and probability who construct normative theories in the first place is, perhaps, unsurprising. Where such evidence converges, we suggest that it is warranted to make claims about whether answers arose through intuitive or analytic thinking, especially if such answers are associated with increased response times. Results of this kind are not always so neat, but we suggest they do warrant theory development and the generation of hypotheses. Moreover, they are also given valuable context by the existence of normative theories. Indeed, in a case where an individual responds with a logically necessary conclusion to a multiple-model syllogism[6], which is produced after an extended period of deliberation, through the consideration of alternative representations and the application of considerable cognitive effort, it would seem unreasonable to judge it as being of equal value to a response produced intuitively, and rapidly that may have involved very little reasoning. From a Meliorist perspective, it is clear that this effortful consideration of multiple models is more desirable than an intuitive, non-logical response and such results provide both an interesting context for normative theories and evidence of further sub-sets of behavior that a descriptivist must account for in their theorizing.

We contend that it is psychologically interesting to investigate reasoners who understand and engage with the experimenter's instructions, reasoners who adopt more nuanced interpretations of quantifiers, and reasoners who actively consider alternative representations and counterexamples – particularly those who do this without formal instruction in the relevant normative theory. We would argue that an abandonment of the research program into the psychological correlates of normative reasoning would be a far more damaging than the potential for theoretical cul-de-sacs that can be generated when philosophical and psychological questions are conflated.

NORMATIVISM AS A SUB-CATEGORY OF INSTRUMENTAL RATIONALITY

An appeal to soft normativism seems to be reflected in Elqayam's (2012) more recent development of a metatheoretical framework that she describes as *grounded rationality*, which involves an extension of her earlier purely descriptivist position (e.g., Elqayam and Evans, 2011). Elqayam's (2012) grounded rationality proposal involves her acceptance of a 'moderate epistemic relativism,' that is, the view that any description of behavior or cognition as rational needs to be grounded by the context in which it takes place. Thus, for example, a slow analytic judgment will always be irrational if it is made too late to be relevant. We agree with Elqayam's (2012) position regarding moderate epistemic relativism, but we take issue with a key aspect of her grounded rationality account, which only allows for a very narrow role for normativism in judging behavior or cognition. The argument is that in order for an inference to be considered as normative the reasoner must adopt the goal of reasoning in accordance with a particular normative theory, with the adoption of such a goal presumably being a *conscious* process. In this way "normative rationality can still be evaluated, albeit as a sub-category of instrumental rationality" (Elqayam, 2012, p. 628). The explicit adoption of a normative theory as an epistemic goal by a reasoner would seem to be an exceptionally rare circumstance. It is far more likely that someone consciously sets out to reason or argue 'rationally' or 'correctly,' but that their knowledge or application of a particular set of normative standards is merely implicit to this goal. Indeed, untrained participants often demonstrate deductive competence when their responses are judged according to logical principles, but this does not mean that their goal was to respond in accordance with a normative system such as logic, nor does it mean that explicit knowledge of logical principles was applied in the production of normatively consistent responses.

Given Elqayam's (2012) apparent proposal that explicit awareness of a normative benchmark is necessary for a reasoning process to be designated as normative – either from a grounded rationality perspective (Elqayam, 2012) or from a $Rationality_2$ perspective (Evans and Over, 1996) – then the attainment of such normativity by a reasoner would only be available to elite participants who have been trained in, for example, formal logic or Bayesian

[6]A multiple-model syllogism is a cognitively demanding reasoning problem where multiple possibilities need to be considered to be certain of what necessarily follows according to formal logic (e.g., see Johnson-Laird and Byrne, 1991).

probability. The untutored will be unlikely to recognize explicitly their analytic thinking as conforming to these criteria and so cannot be described as conforming to $Rational_2$ standards. In fact, we would only really be able to claim that someone has been $Rational_2$ if we asked them after an experiment to tell us which normative standard they were following and they were able to describe this normative standard successfully. Therefore, untutored participants who, during an experiment, set their instrumental goal to follow the instructions, to consider carefully every state of affairs that they can bring to mind and to respond rationally, cannot be considered $Rational_2$ according to Elqayam's proposal. Instead, they would be classified as having produced normative responses via $Rational_1$ processes. The result is, we contend, an incredibly narrow conception of $Rationality_2$, whereby it virtually never occurs in standard reasoning research, where participants are almost always selected because they are naïve to formal logic or some other normative benchmark.

Evans (2007) makes it clear that analytic $Type_2$ thinking is not synonymous with $Rational_2$ thinking. This is not simply because analytic thinking does not always align with normative responding, but because $Type_2$ thinking does not necessarily (or even often) include the conscious goal to reason in accordance with a specific set of normative benchmarks. Moreover, we argue that it should not be claimed that someone is $Irrational_2$ due to their ignorance of normative benchmarks. If someone is responding in the absence of a normative benchmark, rather than contravening a standard that they are aware of, they may be better conceived of as $Arational_2$; only someone who is aware of the appropriate normative theory, but who then fails in their application of it, can be considered to be $Irrational_2$. Participants who avoid the fundamental analytic bias, but are not trained in a particular normative theory are, we argue, very valuable to the development of reasoning theory and are central to the Meliorist agenda. They do not, however, fit neatly into either category of rationality.

While claims that thinking reflects some normative system or that thinking ought to conform to a normative system remain controversial, we argue that thinking can be usefully contrasted with relevant normative systems and that such comparisons inspire and advance the study of the psychology of reasoning. These comparisons should be made with an assumption of bounded rationality (Simon, 1982), that is, with due consideration to the computational demands of tasks and the pragmatic interpretations that people adopt, as well as a realistic stance on the cognitive capacities that we possess. As Stich (1990) famously argued "it seems simply perverse to judge that subjects are doing a bad job of reasoning because they are not using a strategy that requires a brain the size of a blimp" (p. 27). Evans and Elqayam (2011) acknowledge that "paradigms inspired by normativism have led to a number of important psychological findings" (p. 283), and we concur that while these normative theories do not provide perfect foundations for psychological theories of reasoning to be built upon, they do remain a useful benchmark against which to consider participants' reasoning. Furthermore, scrutiny of Meliorist theories from a descriptivist perspective as well as scrutiny of descriptivist theories from a Meliorist perspective has the potential to offer insights for enhancing reasoning and for furthering our ability to describe and understand the cognitive processes that reasoning is underpinned by. In sum, we accept that there are numerous issues with taking an uncritical approach to the use of normative standards in reasoning research, but we also argue that if such standards are discarded altogether we lose the proverbial baby with the bathwater, which undermines our explanations and descriptions of reasoning processes to the point of potential triviality.

REFERENCES

Ackerman, R. (2014). The Diminishing Criterion Model for metacognitive regulation of time investment. *J. Exp. Psychol. Gen.* 143, 1349–1368. doi: 10.1037/a0035098

Ackerman, R., and Thompson, V. A. (2014). "Meta-reasoning: what can we learn from meta-memory," in *Reasoning as Memory*, eds A. Feeney and V. A. Thompson (Hove: Psychology Press).

Ackerman, R., and Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychon. Bull. Rev.* 19, 1189–1192. doi: 10.3758/s13423-012-0305-z

Alter, A. L., and Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Pers. Soc. Psychol. Rev.* 13, 219–235. doi: 10.1177/1088868309341564

Ball, L. J. (2013a). "Eye-tracking and reasoning: what your eyes tell about your inferences," in *New Approaches in Reasoning Research*, eds W. De Neys and M. Osman (Hove: Psychology Press), 51–69.

Ball, L. J. (2013b). Microgenetic evidence for the beneficial effects of feedback and practice on belief bias. *J. Cogn. Psychol.* 25, 183–191. doi: 10.1080/20445911.2013.765856

Ball, L. J., Lucas, E. J., Miles, J. N. V., and Gale, A. G. (2003). Inspection times and the selection task: what do eye-movements reveal about relevance effects? *Q. J. Exp. Psychol.* 56A, 1053–1077. doi: 10.1080/02724980244000729

Ball, L. J., Phillips, P., Wade, C. N., and Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: eye-movement evidence for selective processing models. *Exp. Psychol.* 53, 77–86. doi: 10.1027/1618-3169.53.1.77

Bott, L., and Noveck, I. A. (2004). Some utterances are under informative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006

Buckwalter, W., and Stich, S. (2011). Competence, reflective equilibrium, and dual-system theories. *Behav. Brain Sci.* 34, 251–252. doi: 10.1017/S0140525X11000550

Cherubini, P., Garnham, A., Oakhill, J., and Morley, E. (1998). Can any ostrich fly? Some new data on belief bias in syllogistic reasoning. *Cognition* 69, 179–218. doi: 10.1016/S0010-0277(98)00064-X

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Cohen, L. J. (1981). Can human rationality be demonstrated experimentally? *Behav. Brain Sci.* 4, 317–370. doi: 10.1017/S0140525X00009092

Crupi, V., and Girotto, V. (2014). From is to ought, and back: how normative concerns foster progress in reasoning research. *Front. psychol.* 5:219. doi: 10.3389/fpsyg.2014.00219

De Neys, W. (2006). Automatic-heuristic and executive-analytic processing during reasoning: chronometric and dual-task considerations. *Q. J. Exp. Psychol.* 59, 1070–1100. doi: 10.1080/02724980543000123

De Neys, W. (2012). Bias and conflict: a case for logical intuitions. *Perspect. Psychol. Sci.* 7, 28–38. doi: 10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: some clarifications. *Think. Reason.* 20, 169–187. doi: 10.1080/13546783.2013.854725

De Neys, W., and Bonnefon, J. F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends Cogn. Sci.* 17, 172–178. doi: 10.1016/j.tics.2013.02.001

De Neys, W., Cromheeke, S., and Osman, M. (2011). Biased but in doubt: conflict and decision confidence. *PLoS ONE* 6:e15954. doi: 10.1371/journal.pone.0015954

De Neys, W., and Glumicic, T. (2008). Conflict monitoring in dual process theories of reasoning. *Cognition* 106, 1248–1299. doi: 10.1016/j.cognition.2007.06.002

De Neys, W., Moyens, E., and Vansteenwegen, D. (2010). Feeling we're biased: autonomic arousal and reasoning conflict. *Cogn. Affect. Behav. Neurosci.* 10, 208–216. doi: 10.3758/CABN.10.2.208

De Neys, W., Rossi, S., and Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychon. Bull. Rev.* 20, 269–273. doi: 10.3758/s13423-013-0384-5

Elqayam, S. (2012). Grounded rationality: descriptivism in epistemic context. *Synthese* 189, 39–49. doi: 10.1007/s11229-012-0153-4

Elqayam, S., and Evans, J. St. B. T. (2011). Subtracting 'ought' from 'is': descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Evans, J. St. B. T. (1993). "Bias and rationality," in *Rationality: Psychological and Philosophical Perspectives*, eds K. I. Manktelow and D. E. Over (London: Routledge), 6–30.

Evans, J. St. B. T. (2000). "Thinking and believing," in *Mental Models in Reasoning*, eds J. Garcìa-Madruga, N. Carriedo, and M. J. González-Labra (Madrid: UNED), 41–56.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: extension and evaluation. *Psychon. Bull. Rev.* 13, 378–395. doi: 10.3758/BF03193858

Evans, J. St. B. T. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Hove: Psychology Press.

Evans, J. St. B. T. (2012). "Dual-process theories of deductive reasoning: facts and fallacies," in *The Oxford Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (Oxford: Oxford University Press), 115–133.

Evans, J. St. B. T., and Ball, L. J. (2010). Do people reason on the Wason selection task? A new look at the data of Ball et al. (2003). *Q. J. Exp. Psychol.* 63, 434–441. doi: 10.1080/17470210903398147

Evans, J. S. B., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Mem. Cogn.* 11, 295–306. doi: 10.3758/BF03196976

Evans, J. St. B. T., and Elqayam, S. (2011). Towards a descriptivist psychology of reasoning and decision making. *Behav. Brain Sci.* 34, 275–290. doi: 10.1017/S0140525X11001440

Evans, J. St. B. T., and Lynch, J. S. (1973). Matching bias in the selection task. *Br. J. Psychol.* 64, 391–397. doi: 10.1111/j.2044-8295.1973.tb01365.x

Evans, J. St. B. T., and Over, D. E. (1996). *Rationality and Reasoning*. Hove: Psychology Press.

Evans, J. St. B. T., and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition* 54, 1–71. doi: 10.1016/0010-0277(94)00625-U

Garson, J. W. (2014). "Open futures in the foundations of propositional logic," in *Nuel Belnap on Indeterminism and Free Action*, ed. T. Müller (Heidelberg: Springer International Publishing), 123–145. doi: 10.1007/978-3-319-01754-9_6

Gilhooly, K. J., Logie, R. H., Wetherick, N. E., and Wynn, V. (1993). Working memory and strategies in syllogistic reasoning tasks. *Mem. Cogn.* 21, 115–124. doi: 10.3758/BF03211170

Goel, V., and Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition* 87, B11–B22. doi: 10.1016/S0010-0277(02)00185-3

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goodman, N. (1965). *Fact, Fiction, and Forecast*. Indianapolis, IN: Bobbs-Merrill.

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics 3: Speech Acts*, eds P. Cole and J. L. Morgan (New York, NY: Academic Press), 41–58.

Haigh, M., Stewart, A. J., and Connell, L. (2013). Reasoning as we read: establishing the probability of causal conditionals. *Mem. Cogn.* 41, 152–158. doi: 10.3758/s13421-012-0250-0

Houdé, O. (2007). First insights on "neuropedagogy of reasoning." *Think. Reason.* 13, 81–89. doi: 10.1080/13546780500450599

Johnson-Laird, P. N., and Byrne, R. M. J. (1991). *Deduction*. Hove: Erlbaum.

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Klauer, K. C., and Singmann, H. (2013). Does logic feel good? Testing for intuitive detection of logicality in syllogistic reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1265–1273. doi: 10.1037/a0030530

Kruglanski, A. W., Bélanger, J. J., Chen, X., Köpetz, C., Pierro, A., and Mannetti, L. (2012). The energetics of motivated cognition: a force-field analysis. *Psychol. Rev.* 119, 1–20. doi: 10.1037/a0025488

Lucas, E. J., and Ball, L. J. (2005). Think-aloud protocols and the selection task: evidence for relevance effects and rationalisation processes. *Think. Reason.* 11, 35–66. doi: 10.1080/13546780442000114

Luo, J., Liu, X., Stupple, E. J., Zhang, E., Xiao, X., Jia, L., et al. (2013). Cognitive control in belief-laden reasoning during conclusion processing: an ERP study. *Int. J. Psychol.* 48, 224–231. doi: 10.1080/00207594.2012.677539

Luo, J., Tang, X., Zhang, E., and Stupple, E. J. N. (2014). The neural correlates of belief-bias inhibition: the impact of logic training. *Biol. Psychol.* doi: 10.1016/j.biopsycho.2014.09.010 [Epub ahead of print].

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H. Freeman.

Morley, N. J., Evans, J. St. B. T., and Handley, S. J. (2004). Belief bias and figural bias in syllogistic reasoning. *Q. J. Exp. Psychol. A* 57, 666–692. doi: 10.1080/02724980343000440

Morsanyi, K., and Handley, S. J. (2012). Logic feels so good -I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 596–616. doi: 10.1037/a0026099

Newstead, S. E. (1989). Interpretational errors in syllogistic reasoning. *J. Mem. Lang.* 28, 78–91. doi: 10.1016/0749-596X(89)90029-6

Newstead, S. E., and Griggs, R. A. (1983). Drawing inferences from quantified statements: a study of the square of opposition. *J. Verbal Learning Verbal Behav.* 22, 535–546. doi: 10.1016/S0022-5371(83)90328-6

Nickerson, R. S. (2008). *Aspects of Rationality. Reflections on What it Means to be Rational and Whether We Are*. New York, NY: Psychology Press.

Noveck, I. A., and Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends Cogn. Sci.* 12, 425–431. doi: 10.1016/j.tics.2008.07.009

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 608. doi: 10.1037/0033-295X.101.4.608

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198524496.001.0001

Prowse Turner, J. A., and Thompson, V. A. (2009). The role of training, alternative models and logical necessity in determining confidence in syllogistic reasoning. *Think. Reason.* 15, 69–100. doi: 10.1080/13546780802619248

Quayle, J. D., and Ball, L. J. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. *Q. J. Exp. Psychol.* 53A, 1202–1223. doi: 10.1080/713755945

Rips, L. J. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA: MIT Press.

Roberts, M. J. (1993). Human reasoning: deduction rules or mental models, or both? *Q. J. Exp. Psychol.* 46, 569–589. doi: 10.1080/14640749308401028

Roberts, M. J., Newstead, S. E., and Griggs, R. A. (2001). Quantifier interpretation and syllogistic reasoning. *Think. Reason.* 7, 173–204. doi: 10.1080/13546780143000008

Samuels, R., and Stich, S. P. (2004). "Rationality and psychology," in *The Oxford Handbook of Rationality*, eds A. R. Mele and P. Rawling (Oxford: Oxford University Press), 279–300.

Schmidt, J. R., and Thompson, V. A. (2008). 'At least one' problem with 'some' formal reasoning paradigms. *Mem. Cogn.* 36, 217–229. doi: 10.3758/MC.36.1.217

Shynkaruk, J. M., and Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Mem. Cogn.* 34, 619–632. doi: 10.3758/BF03193584

Simon, H. A. (1982). *Models of Bounded Rationality: Empirically Grounded Economic Reason*, Vol. 3. Cambridge, MA: MIT Press.

Singmann, H., Klauer, K. C., and Kellen, D. (2014). Intuitive logic revisited: new data and a Bayesian mixed model meta-analysis. *PLoS ONE* 9:e94223. doi: 10.1371/journal.pone.0094223

Smedslund, J. (1990). A critique of Tversky and Kahneman's distinction between fallacy and misunderstanding. *Scand. J. Psychol.* 31, 110–120. doi: 10.1111/j.1467-9450.1990.tb00822.x

Stanovich, K. E. (1999). *Who is Rational?: Studies of Individual Differences in Reasoning*. Mahwah, NJ: Psychology Press.

Stanovich, K. E. (2009). "Distinguishing the reflective, algorithmic and autonomous minds: is it time for a tri-process theory?," in *In Two Minds: Dual Processes and*

Beyond, eds J. St. B. T. Evans and K. Frankish (Oxford: Oxford University Press), 55–88.

Stanovich, K. E. (2011). Normative models in psychology are here to stay. *Behav. Brain Sci.* 34, 268–269. doi: 10.1017/S0140525X11000161

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* 23, 645–665. doi: 10.1017/S0140525X00003435

Stanovich, K. E., West, R. F., and Toplak, M. E. (2010). "Individual differences as essential components of heuristics and biases research," in *The Science of Reason: A Festschrift for Jonathan St. B. T. Evans*, eds K. Manktelow, D. Over, and S. Elqayam. (Hove: Psychology Press), 355–396.

Stenning, K., and Cox, R. (2006). Reconnecting interpretation to reasoning through individual differences. *Q. J. Exp. Psychol.* 59, 1454–1483. doi: 10.1080/17470210500198759

Stewart, A. J., Haigh, M., and Ferguson, H. J. (2013). Sensitivity to speaker control in the online comprehension of conditional tips and promises: an eye-tracking study. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1022–1036. doi: 10.1037/a0031513

Stich, S. P. (1990). *The Fragmentation of Reason.* Cambridge, MA: MIT Press.

Stich, S. P., and Nisbett, R. E. (1980). Justification and the psychology of human reasoning. *Philos. Sci* 47, 188–202. doi: 10.1086/288928

Stupple, E. J. N., and Ball, L. J. (2007). Figural effects in a syllogistic evaluation paradigm: an inspection-time analysis. *Exp. Psychol.* 54, 120–127. doi: 10.1027/1618-3169.54.2.120

Stupple, E. J. N., and Ball, L. J. (2008). Belief–logic conflict resolution in syllogistic reasoning: inspection-time evidence for a parallel process model. *Think. Reason.* 14, 168–189. doi: 10.1080/13546780701739782

Stupple, E. J. N., and Ball, L. J. (2011). Normative benchmarks are useful for studying individual differences in reasoning. *Behav. Brain Sci.* 34, 270–271. doi: 10.1017/S0140525X11000562

Stupple, E. J. N., Ball, L. J., and Ellis, D. (2013). Matching bias in syllogistic reasoning: evidence for a dual-process account from response times and confidence ratings. *Think. Reason.* 19, 54–77. doi: 10.1080/13546783.2012.735622

Stupple, E. J. N., Ball, L. J., Evans, J. S. B. T., and Kamal-Smith, E. (2011). When logic and belief collide: individual differences in reasoning times support a selective processing model. *J. Cogn. Psychol.* 23, 931–941. doi: 10.1080/20445911.2011.589381

Stupple, E. J. N., and Waterhouse, E. F. (2009). Negations in syllogistic reasoning: evidence for a heuristic-analytic conflict. *Q. J. Exp. Psychol.* 62, 1533–1541. doi: 10.1080/17470210902785674

Thompson, V. A. (2009). "Dual-process theories: a metacognitive perspective," in *In Two Minds: Dual Processes and Beyond*, eds J. Evans and K. Frankish (Oxford: Oxford University Press), 171–195.

Thompson, V. A. (2011). Normativism versus mechanism. *Behav. Brain Sci.* 34, 272–273. doi: 10.1017/S0140525X11000574

Thompson, V. A., Morley, N. J., and Newstead, S. E. (2011a). "Methodological and theoretical issues in belief-bias: implications for dual process theories," in *The Science of Reason: A Festschrift for Jonathan St. B. T Evans*, eds K. I. Manktelow, D. E. Over, and S. Elqayam (Hove: Psychology Press), 309–338.

Thompson, V. A., Prowse-Turner, J., and Pennycook, G. (2011b). Intuition, reason, and metacognition. *Cogn. Psychol.* 63, 107–140. doi: 10.1016/j.cogpsych.2011.06.001

Thompson, V. A., Prowse-Turner, J., Pennycook, G. R., Ball, L. J., Brack, H. M., Ophir, Y., et al. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition* 128, 237–251. doi: 10.1016/j.cognition.2012.09.012

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., and Campbell, J. D. (2003). Syllogistic reasoning time: disconfirmation disconfirmed. *Psychon. Bull. Rev.* 10, 184–189. doi: 10.3758/BF03196483

Topolinski, S. (2014). "Intuition: introducing affect into cognition," in *Reasoning as Memory*, eds A. Feeney and V. A. Thompson (Hove: Psychology Press).

Topolinski, S., and Reber, R. (2010). Gaining insight into the "aha" experience. *Curr. Dir. Psychol. Sci.* 19, 402–405. doi: 10.1177/0963721410388803

Trippas, D., Handley, S. J., and Verde, M. F. (2013). The SDT model of belief bias: complexity, time, and cognitive ability mediate the effects of believability. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1393–1402. doi: 10.1037/a0032398

Trippas, D., Handley, S. J., and Verde, M. F. (2014). Fluency and belief bias in deductive reasoning: new indices for old effects. *Front. Psychol.* 5:631. doi: 10.3389/fpsyg.2014.00631

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Wason, P. (1966). "Reasoning," in *New Horizons in Psychology*, ed. B. Foss. (Middlesex: Penguin Books), 135–151.

Weston, A. (2009). *A Rulebook for Arguments.* Indianapolis, IN: Hackett Publishing.

Wetherick, N. E., and Gilhooly, K. J. (1995). 'Atmosphere,' matching, and logic in syllogistic reasoning. *Curr. Psychol.* 14, 169–178. doi: 10.1007/BF02686906

Wilkinson, M. R., Ball, L. J., and Cooper, R. (2010). "Arbitrating between Theory–Theory and Simulation Theory: evidence from a think-aloud study of counterfactual reasoning," in *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society*, eds S. Ohlsson and R. Catrambone (Austin, TX: Cognitive Science Society), 1008–1013.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 February 2014; accepted: 19 October 2014; published online: 05 November 2014.

Citation: Stupple EJN and Ball LJ (2014) The intersection between Descriptivism and Meliorism in reasoning research: further proposals in support of 'soft normativism.' Front. Psychol. ***5****:1269. doi: 10.3389/fpsyg.2014.01269*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

GLOSSARY

ARATIONAL
Neither rational nor irrational, but instead existing outside of the category of rationality.

BELIEF BIAS
The tendency to judge the validity of an argument based on the believability of its conclusion rather than on whether the conclusion is logically necessitated by the argument's premises.

DESCRIPTIVISM
The view that normative standards are not appropriate benchmarks in cognitive science and that the goal of psychological research is to *describe* behavior without making value judgments.

DOUBLE NEGATION ELIMINATION
The inference that if not not-A is true then A is true (and its converse), which is proposed by Rips (1994) as a simple, intuitive logical rule.

MELIORISM
In general usage, Meliorism is the belief that humans can improve the world. In the present context the term is used specifically to refer to the idea that thinking, reasoning and judgment can be enhanced through education, training and practice. Meliorism in this latter sense also reflects a research program in Cognitive Science.

NORMATIVE
Refers to the 'correct' answer or the 'right' way of doing things. In the present context, normative benchmarks are the (often debatable) standards for thinking, reasoning or deciding that participant responses tend to be evaluated against. These normative benchmarks derive from formal, logical systems or probability theory.

PANGLOSSIAN
Derived from Dr. Pangloss, the eternal optimist in Voltaire's *Candide*, Panglossian refers to the belief that 'all is for the best in the best of all possible worlds.' In the present context, it is the idea that we have the best of all possible cognitive systems.

RATIONALITY$_1$
Thinking, speaking, reasoning, making a decision, or acting in a way that is generally reliable and efficient for achieving one's goals.

RATIONALITY$_2$
Thinking, speaking, reasoning, making a decision, or acting when one has a reason for what one does sanctioned by a normative theory.

SLIPPERY SLOPE FALLACY
The argument that a relatively small first step leads inevitably to the bottom of the slippery slope, so if A happens then B will happen and if B happens then C will happen, all the way down to the terrible scenario of Z.

frontiers in PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 20 October 2014
doi: 10.3389/fpsyg.2014.01159

The empirical study of norms is just what we are missing

Theodora Achourioti[1], Andrew J. B. Fugard[2]* and Keith Stenning[3]

[1] *Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands*
[2] *Department of Clinical, Educational and Health Psychology, University College London, London, UK*
[3] *Department of Psychology, University of Gießen, Giessen, Germany*

Edited by:
David E. Over, Durham University, UK

Reviewed by:
Shira Elqayam, De Montfort University, UK
Catarina Dutilh Novaes, University of Groningen, Netherlands

****Correspondence:***
Andrew J. B. Fugard, Research Department of Clinical, Educational and Health Psychology, University College London, 26 Bedford Way, London WC1H 0AP, UK
e-mail: a.fugard@ucl.ac.uk

This paper argues that the goals people have when reasoning determine their own norms of reasoning. A radical descriptivism which avoids norms never worked for any science; nor can it work for the psychology of reasoning. Norms as we understand them are illustrated with examples from categorical syllogistic reasoning and the "new paradigm" of subjective probabilities. We argue that many formal systems are required for psychology: classical logic, non-monotonic logics, probability logics, relevance logic, and others. One of the hardest challenges is working out what goals reasoners have and choosing and tailoring the appropriate logics to model the norms those goals imply.

Keywords: reasoning goals, normativity, reasoning norms, syllogisms, classical logic, nonmonotonic reasoning, probabilistic reasoning, heterogeneity of human reasoning

1. INTRODUCTION

Formal systems offer a precise way to characterize people's various reasoning goals. There are many logics for different situations. Some allow reasoners to withdraw conclusions as more information is learned. Others describe the logic of deontic rules about "ought" and "must." There are logics for relevance and for probabilities. Each logic provides different norms, e.g., for what constitutes a valid logical argument or whether a sentence is true. Elqayam and Evans (2011) propose that normativity in psychological practice should be avoided. We cannot see how. In this article we argue that without norms of some kind, we cannot interpret the data participants produce. Rather, participants' reasoning goals generate their own norms of reasoning and logics provide a good way to capture these norms. Pure descriptivism is impossible, and highly undesirable.

We first remind the reader of the distinction between *constitutive* and *regulative* norms which plays an important role in this paper. Constitutive norms define a certain behavior for what it is (see Searle, 1969). Characteristic examples are the rules of a game, e.g., the game of chess: changing the rules means playing a different game. Norms are regulative rather than constitutive when they do not define but regulate a preexisting activity. In this sense, regulative norms are not necessary and they are also derivative: they are consequences of constitutive norms, together with contextual features such as overall goals or specific constraints. For instance, what move to perform at any point when playing a game of chess is dictated by regulative norms: it may be that one wants to lose and terminate the game as soon as possible. Even with this unusual contextual goal, the revised regulative norms arise from the usual constitutive norms. Importantly, regulative norms are action oriented, in the sense that they tell one what to do.

Formal systems are instrumental in specifying constitutive and regulative norms, which is in turn necessary in order to understand what participants do in a particular reasoning task. Formal systems are characterized by constitutive norms: doing arithmetic is constituted by complying with the well known constitutive norms of arithmetic. And constitutive norms give rise to regulative norms (Achourioti et al., 2011). If you are dealing with numbers that represent prices of items, and you want a total, then adding them is permissible—a regulative norm. If you are dealing with numbers which are barcode identifiers and you want to count tokens (stocktaking perhaps?), then adding two of them is nonsense—another regulative norm. Formal systems impose regulative norms on non-formal activities that use them, and they do it as a consequence of their constitutive norms. Not uniquely of course, as our examples of trying to lose at chess, and different activities with numbers show. What the regulative norm is depends on the goals and other contextual features at hand; and as goals may be radically different (think of our earlier example of someone playing chess to lose), the regulative norms they generate may be radically different too.

Norms and values are, in the crucial cases for the psychology of reasoning, the least observable features of thinking—the farthest from being fixed by data without system or theory. Participants generally cannot describe their goals in the terms of appropriate systems or theory. Their performances nevertheless can provide evidence for theory-relative normative specification of goals, once a formal analysis is available. In this paper we illustrate these points with experimental examples.

There certainly are abuses of norms to be observed. We propose that these are most evident when any single homogeneous system account of human reasoning is proposed, whether it be classical logic (CL), probability theory, or indeed radical descriptivism with a single description language. As soon as hegemony is proposed, it becomes impossible to study the basis for selection from among multiple systems of reasoning, and it is this

requirement to select from multiple possible systems that most clearly dissolves perceived problems of normativity, and connects reasoning goals to instrumental goals. Selecting from multiple possible reasoning goals can be done on instrumental grounds suiting the goals to the problem at hand. We do not believe there is any such thing as "human reasoning" construed as a homogenous system for the simple reason that the demands of different reasoning problems are incompatible, as we illustrate below. The main reasoning goal of this paper it to illustrate this point with examples from past and current practice.

The backdrop to our approach to norms and normativity is the multiple-logics approach to human reasoning outlined in Stenning and van Lambalgen (2008). It is widely accepted in modern logic that there are many logics which capture many kinds of reasoning, often incompatible one with another. They are best thought of as mathematical models of pure archetypes of reasoning. Logics have been around for a while, however, with notable exceptions, psychology still mostly uses only classical ("textbook" logic) and probability logics, and often rejects the idea that the latter even is a logic. What goes for logics goes more generally for formal systems used for modeling cognition. We therefore begin by providing some triangulation points better known to psychologists that relate this framework to possibly more familiar territory.

Todd et al. (2012) have proposed a multiple heuristics approach to decision making which makes the choice of alternative methods a contextualized choice, and in this shares important features with our multiple-systems approach to reasoning. The resulting norms are content-dependent as argued by Gigerenzer (2001). Bayesian models are often viewed as the established norm in decision, as well as more recently in reasoning. Todd et al. (2012) argue against the universality of a probabilistic norm. The heuristics proposed are specialized, and logics are at a somewhat different level of analysis, so not easy to compare, but nevertheless the two approaches are more closely related than may at first appear. Existing neural networks which implement the nonmonotonic logic we use, Logic Programming (LP) (Stenning and van Lambalgen, 2008, chapter 7), along with the internal generation of statistics of the networks' operation, can supply the theory-relative conditional frequency information that is required to select for these heuristics the content that they require in context. The networks also provide lists of defeaters—conditions that defeat conditional inferences and contribute to determining confidence in causal conditional reasoning (Cummins, 1995). This therefore offers a qualitative system of graded uncertainty in intensional reasoning which is a competitor to Bayesian methods in some contexts, through implementing the decision heuristics just mentioned.

Stich (1990) "The Fragmentation of Reason" and this author's work more generally on cognitive pluralism, is chiefly focussed on cases where different people (or peoples) have different norms of reasoning for some reason of individual or cultural preference or habit. We are focussed on cases in which participants' various goals call for different logics or systems of reasoning in different contexts. At least at first pass, on our account, everyone ought to conform to the constitutive norms of classical logic if their goals are, say, classical mathematical proof or the settlement of a certain kind of dispute. Everyone ought to conform to the norms of some nonmonotonic logic such as LP if their goal is to tell a story. Everyone ought to conform to the norms of deontic logic if they want to reason about permissions and obligations. And so on. So, our proposal is not relativistic in the usual sense. It is relativistic only in the sense that people's goals and therefore their norms are variable in different contexts. This does not diminish the interest of Stich's topic, nor of the two topics' relatedness. Widlok and Stenning (submitted) sketch how a multiple-logics approach bears on the recurrent anthropological debate about whether different cultures have different logics. Using nonmonotonic LP to analyse the Mambila's discourse of divination by spider, it concludes that cultures vary in the social circumstances in which they bring logics to bear, but that a working hypothesis should be that they evidence the same range of logics in the range of contexts they experience. Spider divination in context looks a whole lot less irrational through these eyes.

Clearly many authors have proposed many heterogeneities in reasoning, such as what is conventionally meant by the phrase "individual differences" in psychology, individual variation in how "good" some performance is. We are here concerned with a specific type of (in)homogeneity of formal system (e.g., classical logic, probability, nonmonotonic logic, ...). Elqayam (2012) proposes grounded rationality—essentially the avowedly uncontroversial proposal that there is more to rational reasoning and action than the adoption of a formal system. There is more because people differ in their cognitive capacities, cognitive costs, mundane aims, and all the other variables of bounded rationality, and more. Elqayam (2012) appears to associate normativism with the adoption of a single formal standard of reasoning (usually either classical logic or probability in some form), and proposes "descriptivism" as an alternative that can preserve variety. So we agree there is more to rational action than logics or formal systems, and that adoption of a single system is a mistake. But we disagree that "descriptivism" can be conceived as an alternative to multiple-systems, and propose that the mundane limitations of grounded and bounded rationality interact with the unavoidable choice of reasoning system among the other systems that are also required. It is this interaction that provides great opportunity and power to the empirical investigation of reasoning and rationality. Description is of course important, but is always theory- and goal-relative. Since there are many theories and goals, there are many descriptions, and description itself cannot solve the inevitable choice of interpretation problem.

Bounded rationality is a proposal (which we applaud) that rational action has to be understood as governed by the intersection of many systematic constraints. To take one of Simon's examples (Simon, 1972), if working memory limitations are an important bound on a particular reasoning task, then a theory of working memory will be required to intersect with the cognitive implementation of whatever reasoning system is at work, in order to understand how contextual features (whether we have pencil and paper, whether we are expert in the domain, ...) affect performance, and therefore what constitutes rational action for us in context. Countless social bounds are also sources of systematic constraint. Many relevant features of any particular situation may be entirely due to coincidence, but their operation is nevertheless

to be understood in terms of several systematic theories. Totally unsystematic constraints are not comprehensible, by hypothesis. Thus bounded (or grounded) rationality requires multiple simultaneous systematic formal accounts of all the relevant constraints. With these systems come constitutive norms; and with those constitutive norms come regulative norms. The fact that we are not currently in a position to specify the many systematic constraints in general terms, and that we can make some progress with rather *ad hoc* accounts of say working memory, does not make a theory of bounded rationality able to dispense with these intersecting generalizations[1]. Boundedness does not make rationality *ad hoc*. The boundedness of working memory may or may not be there because we *ought* to be bounded in memory (though see, for example, Hertwig and Todd, 2005 and MacGregor, 1987 where advantages of boundedness are proposed) but it generates regulative norms such as: for an important reasoning task that clearly overloads your unaided working memory, it is not rational, other things being equal, not to have a pencil and paper to hand. Although we deliberately use examples of norms arising from individual reasoning because they are how experimental psychology generally meets up with normative considerations, it is not hard to see that the regulative norms arising from the constitutive norms of the formal elements can rapidly reach into any social, ethical or political activity people engage in.

As yet another orientation point, we recall that more than one logic may operate within an activity. Elsewhere we have proposed that an account of how at least some kinds of argument work, requires an account of how adversarial classical and cooperative nonmonotonic logics have to work together (Stenning, 2002, chapter 5, Stenning, 2012) to capture the interplay between cooperative and adversarial relations in argument. Mercier and Sperber (2011) propose that reasoning evolved for argumentation. These authors define reasoning with respect to explicitly aware processes, relegating unconscious processes to mere "inference." On our account, accounting for argumentation that calls on both non monotonic and monotonic logics means bridging what Mercier and Sperber divide between inference and reasoning. One might propose that once cooperative discourse became possible, argumentation about its interpretation inevitably followed, for monitoring and repairing breakdowns in understanding. Argumentation is inconceivable without the existence of cooperative discourse. Elsewhere, we have criticized adaptationist attempts to try to read evolutionary accounts from informal descriptions of current function (Stenning and van Lambalgen, 2008, chapter 6). What is first required is a deeper description of the phenotype: and that requires empirical description of goals and norms.

The plan of this paper is that the first section discusses norms as we understand them, and how they are incompatible with any pure descriptivism. We will concentrate on how participants' very own reasoning goals create variety in *internal* norms which need to be captured in logics before any data of reasoning becomes interpretable, and draw out some consequences for empirical research. If normativity itself is not the problem, it is not without its abuses. We see the homogeneous application of formal systems as a major problem. Once only one system is allowed (whether it is Bayesianism, or classical logic, or whatever) then there is no way of assessing why a system is an appropriate choice for modeling an instance of reasoning. It cannot be an appropriate choice because it is no longer a choice. If there is heterogeneity (many logics or other competence models) then there have to be criteria of application, and indeed choice can be made on instrumental grounds—that is by a match between logical properties and reasoning goals, as we illustrate.

The second section takes the psychological study of categorical syllogistic reasoning as an example to illustrate these points. It argues that the descriptivism prevailing for the last half of the 20th century was exactly what led to a catastrophic inattention to the participants' reasoning goals. It describes the pervasive ambiguity of reasoning experiments for participants, most of whom adopt nonmonotonic reasoning goals where experimenters assumed classical logical ones. It spells out how the contrasting reasoning goals are constituted in the properties of these two logics.

The distinctive properties of classical logic give guidance for design of a context which should improve the chances that we see classical reasoning—in this case a context of dispute. Some results from an ongoing experimental program show how the properties of classical logic which make it suitable for a model of a certain kind of dispute or demonstration are presented as a first indication of the rewards of this kind of empirical program. It provides clear evidence that this context produces more classical reasoning than the conventional draw-a-conclusion task. And perhaps more importantly, it shows how participants have surprising implicit knowledge of some of the peculiarities of classical logic. Psychologically, our goal should be assessing peoples' implicit knowledge and its contextual expression i.e., their implicit logical concepts, rather than their scores on some fixed-context arbitrary task which engenders variable and unspecified goals.

The third section pursues similar themes in the example of probabilistic reasoning. The idea that Bayesianism, or even probability, provides a new homogeneous norm for human reasoning, and for rational action in general, has supplanted the same role that was previously assigned to classical logic in theories of rationality. But probability theory fails to provide reasoning goals at levels comparable to the examples of the previous section. What is argued for is an analogous differentiation of "probability logics" to apply to different reasoning goals, bridging to neighboring logics in a friendly welcoming manner.

Finally we end with some conclusions about the empirical programs that should follow from our arguments for a multiple-logics view of human reasoning, based on the differentiated reasoning goals that this multiplicity affords, together with some comments about the very different view of the relation between logic and psychology which emerges.

[1]For example, one of the prominent accounts of long-term/working-memory interactions (Anderson, 1983) contains a production system which is a specific implementation of LP, the nonmonotonic logic we employ here. So logic is also not so distant from the WM component of bounded rationality. Many psychologists regard retrieval of relevant information from long-term memory as memory rather than reasoning. It is certainly memory, but equally certainly reasoning.

2. EXPLAINING NORMATIVITY

The experimental work discussed in the next two sections is intended to emphasis the role of normativity in the psychology of reasoning and should be read as such. It becomes for this reason important that we clarify what we mean by "normativity" and we will do this by reference to Elqayam and Evans (2011) which argues for descriptive as opposed to normative approaches and encapsulates our main focus. This article was followed by a series of commentaries some of which present views that are close to the points we make here. But we find that in many cases the picture is rather blurred and clarification of the key concepts is much needed so that points of agreement or disagreement can be identified and an essential discussion on the foundations of psychology of reasoning can get off the ground. Importantly, many of the arguments put forward against the use of normative frameworks depend on a specific understanding of "normativity," which we would like to challenge[2].

Logic is often said to be a normative system contrasted with descriptive frameworks that psychologists use. But a logical framework in itself is not descriptive or normative; it is the *use* of a logic that can be descriptive or normative, and even classical logic can serve as a descriptive tool in situations where people are found to reason classically. As we discuss later, such situations do not only arise in specialized contexts such as mathematical reasoning but may be found in research areas as prominent as syllogism tasks or natural language conditional statements. The interesting, indeed normative, question then is what are the circumstances, if there are any, that trigger classical reasoning, and make it appropriate in the situation: when is CL adopted by the participant as their norm for the task? We will discuss how classical logic, and especially those characteristics of it that distinguish it from other formal frameworks, provide cues as to where to look for the goals that may make it appropriate. The same goes for any other logic or formal system.

The role of normativity in questions such as the one just stated is clearly not of the evaluative kind. Contrast this with the following:

> "A normative theory asks evaluative 'ought' questions: 'What *ought* to be the good use of negation in language?' A normative approach contains an element of evaluation, a sense of 'goodness' and 'badness', or 'right' and 'wrong', that is absent from a purely competence account. In short, normative theories are 'ought'-type theories; computational theories are 'is'-type theories. Note that the competence theories and performance theories are both descriptive—what they share is the *is*. " (Elqayam and Evans, 2011), p.239

Here the term "normative" takes on almost ethical connotations. To be sure, such questions of prescriptive "goodness" and "badness" are at best outdated and in any case certainly irrelevant to the study of human reasoning. Not so, however, for "right" and "wrong" questions, as witnessed, for example, when participants report "errors" in their own reasoning and correct themselves in the process (we see an example later in how people reason about uncertain conditionals). There is nothing ethically objectionable or evaluative to supposing that humans are not perfect thinking machines and sometimes commit errors or refrain from driving their reasoning all the way to its utmost consequences[3]. and the notion of "error" makes little sense outside a normative framework that specifies what counts as "right" inferencing and what as "wrong." The pertinent question is rather: how can we talk about "correctness," or "right" and "wrong," without falling into the same old trap as when psychologists considered classical logic to be the arbitrer of human rationality?

Most of the reluctance to engage seriously with normative considerations comes from an understanding of norms as "external" to one's reasoning, that is, as set by someone other than the participant herself (often researchers). Objections to normativity disappear as soon as attention shifts to norms that are constitutive of one's own reasoning, meaning that they help define reasoning for what it is[4]. We do not deny that norms 'set by other people' (social norms) are important. But if it is only such norms that are objectionable the debate has been ill-specified, and the objections to norms should be suitably diluted. A way to trace "internal" norms is to identify the goals that underlie and drive one's reasoning process. Goals are highly complex and not easy to specify as they stem from various sources. They are not observable and they interact with each other in complicated ways. In reasoning experiments, for example, the participant has to decide how to go about solving the task, which depends on the participant's interpretation of what is asked of her, which in turn depends on pragmatic goals influencing natural language processing of instructions, how much is underdetermined by the experimenter's design and so on. But whatever the underlying goals turn out to be, it has to be recognized that they heavily influence the type of reasoning participants engage in. In the next section we discuss concrete examples of how different goals trigger different reasoning processes, and we show this by varying the context in order to generate different types of reasoning (and thereby different reasoning norms) and study the effects of this variation on the experimental data.

With the understanding of normativity that we propose as "internal" and not "external" to reasoning, the discussion of human rationality can be set on new grounds. Consider the following:

> What seems to set apart normative rationality from other types of rationality is the "oughtness" involved in normativism. Bounded rationality, for example, is not bounded because it "ought" to be so. Instead, there are just biological limits to how large brains can

[2]For what Elqayam and Evans (2011) argue against, the term "normativism" seems to us more appropriate than "normativity." This is indeed the term that these authors use, while many of the commentators talk about "normativity." This is not to say that the differences of opinion are merely terminological; it is rather the choice of key terms that is influenced by the theoretical positions adopted.

[3]The authors seem to take issue with the concept of "error" because it evidences the use of norms: 'While the term "normative" has been dropped, the term "error" has not: A recent book (Stanovich, 2009) presents an extensive discussion of the source of reasoning and decision-making errors, implying norms'. (Elqayam and Evans, 2011), p.242

[4]We discuss constitutive and regulative norms and their relations also in Achourioti et al. (2011).

grow and how much information and how many computational algorithms they can store and execute.' (Elqayam and Evans, 2011), p. 236

As mentioned above, even this is contentious in the literature: there may be distinct advantages to limited systems, and there is much evidence that human brain-size is under selective pressure from both directions. But we accept that resource bounds are a fact. Resource constraints certainly influence the reasoning that participants engage in; this is one of the reasons that may render classical model theoretic thinking intractable and force naive participants to resort to nonmonotonic example construction through preferred models, that leads to more manageable computational processes. But notice that participants are switching reasoning subgoals, not attempting the same goal with a different tool. Such limitations are part of what a formal model helps represent. They lie, for example, at the heart of the difference between monotonic and nonmonotonic systems. And justifying one model rather than another is clear evidence of normative status, even if the norms in this case could not be otherwise because of resource bounds. Elqayam and Evans (2011) follow Evans and Over (1996) in setting apart "normative" rationality from "instrumental," "bounded," "ecological" and "evolutionary" rationality. The way we understand normativity, it is integral part of all of these four types of rationality. In fact, most of the present paper discusses norms that are part of so-called "instrumental rationality." Hence, we take issue with remarks as the following:

'Some researchers have proposed that we should adopt alternative normative systems such as those based on information, probability, or decision theory (Oaksford and Chater, 1991, 1998a,b, 2007), while others proposed that at least some forms of rationality need not necessarily require a normative system at all (e.g. Evans, 1993, 2002; Evans and Over, 1996; Gigerenzer and Selten, 2001). By this position, organisms are rational if they act in such a manner as to achieve personal goals, and such rationality need not involve any normative rule following.' (Elqayam and Evans, 2011), p.234

The message here is that achieving personal goals need not involve normative rule following. It must be clear by now that we take reasoning goals to be intrinsically normative in that they play a big role in the choice of one reasoning mode rather than another (without claiming that some conscious decision-making process of selection takes place, or that they are necessarily constituted as such in "rules"). Pragmatic goals of relevance, for example, are essentially normative when in some contexts they exclude the interpretation of a natural language "or" as the classical logic disjunction, $\vee$. Just as with the selection task, examination has to reveal these hidden normative systems behind, for example, ecological rationality. Martignon and Krauss (2003) argue that Gigerenzer's heuristics require Bayesian methods for their population with content in context. And Martignon et al. (in preparation) give an account of this same process based on nonmonotonic logic. Ecological rationality is up to its ears in normativity.

We have so far proposed an understanding of normativity as applying to the use of formal systems rather than attaching to the systems themselves and as involving questions of correctness that do not have evaluative connotations but refer to norms which are internal to human reasoning and constitutive of it. To clarify these points even further, we now discuss the status of competence theories and the "is-ought" fallacy which normative approaches are said to commit. Here is an interesting quote:

'. . . arbitrating between competing normative systems is both crucial and far from easy. This is where the difference between normative and competence theories becomes critical. Competence theories are descriptive and can hence be supported by descriptive evidence. In contrast, can one support normative theory with descriptive evidence? Can one infer the *ought* from the *is*?' (p. 240)

We do not agree that competence theories can be supported by descriptive evidence without normative considerations. It is especially competence theories that have to see beyond the data in order to account for the discrepancy between theory and observation. And at the same time it is a truism that the further one moves away from observable data the more difficult it becomes to actually test the theory. So how is it possible at once to model competence and stay as close as possible to actual performance? Competence theories have constitutive norms, and these norms generate regulative norms once their reasoning is embedded in action. Our examples in the next sections show how the various constitutive norms participants adopt for syllogistic and probabilistic reasoning (competence theories) generate regulative norms once embedded in actual reasoning. A proper understanding of the data depends on the choice of logical norm.

Elqayam and Evans (2011) argue that much of the experimental cognitive research is liable to the "is-ought" fallacy (or naturalistic fallacy as it is often called by philosophers). However, in order for this transition from "is" to "ought" to make sense, "is" and "ought" must be clearly separated, and we show in this paper that descriptive and normative matters cannot be so neatly set apart. A purely descriptive approach is simply unattainable, since what the participants "do" already depends on the theoretical framework within which one performs the observation and this theoretical framework must take into account the reasoning goals at hand, the latter clearly creating normative demands. The dependence of description on formal theory is clearly seen when incompatible descriptions match the same data; when, as we discuss, for instance, the same answer to a reasoning task could be generated by reasoning processes that are as different as monotonic and nonmonotonic logics.

Interestingly, Elqayam and Evans (2011) take the "is-ought" fallacy to be especially triggered in cases where more than one theory matches the data, which then lends support to descriptive theories in their approach[5]. But we believe that it is precisely the need to select among equally matching theories that proves descriptivism to be impossible, on the one hand, and what saves the psychologist from the homogeneity trap, on the other. There we think, is the real danger when studying human reasoning without making explicit the norms and goals involved; namely, the

[5]It must be clear by now that we do not subscribe to a distinction of formal systems into normative and descriptive; it is rather the use we put these systems to in accounting for human reasoning that can be labeled as such.

idea that a single theory can play the role of setting the basis, descriptive or normative, over which to design and assess all experimental work.

Having to arbitrate between formal models is not in itself a problem we should want to eliminate, but it becomes such a problem if it means having to choose between theories that claim to explain human reasoning as a whole. This is where a multiple-logics approach as advocated here offers an improvement in the way formal models are used: in order to account for differences between participants' reasoning within a particular task, we ask ourselves how we can modify the task so that these differences become apparent. This we find the most interesting experimental challenge, which relies, however, on being open to different formalizations sensitive to participants' underlying norms and goals. Formalizing involves representation of reasoning norms (which are goal-sensitive) as much as empirical engagement. And here is where a single descriptive framework, even if that were possible, is bound to fail: it offers no way to account for pervasive participant differences flowing from different goals, if all one is allowed to do is to "describe" participants' micro-behavior.

3. THE SYLLOGISM AS ILLUSTRATION

3.1. REASONING GOALS AS NORMS EMBODIED IN FORMAL SYSTEMS

The earliest paper on the psychology of the syllogism by Störring (1908) does not address the relation between logic and psychology at all, but employing great logical and psychological insight gets on with describing a small number of participants' responses to syllogistic problems. It identifies Aristotle's *ekthesis* as a good guide to participants' reasoning processes. This itself is remarkable, coming so soon after the "divorce" of logic and psychology, and the establishment of the latter as experimental science. By mid-century, Wason (1968) argues strongly against the very idea that logic bears any useful relation to human reasoning, claiming to demonstrate this fact experimentally with Piaget's theory as his target.

It was a further half century before Wason's interpretation of his experiment was prominently challenged in psychology (Chater and Oaksford, 1999; Stenning and van Lambalgen, 2001; Evans, 2002; Stenning and van Lambalgen, 2004) (but see also Wetherick, 1970) by showing how it rested on the assumption that classical logic had to be the goal of participants' supposedly failed reasoning in Wason's Task, for any of his arguments for irrationality to succeed. But it behooves someone so vehement that logic contributes nothing to understanding human reasoning to perhaps find out what constitutes a logic. This simultaneous coupling of explicit denial of the relevance of classical logic, with its under-the-counter adoption as *the* criterion of correct reasoning, stems directly from an avoidance of the issue of participants' goals in reasoning, and this in turn is a direct result of the suppression of formal specifications of reasoning goals, in favor of a proposed descriptivism treating "human reasoning" as an activity with a homogeneous goal. Wherever descriptivism is espoused we find tacit appeal to homogenous normativism.

As we shall see in our example of the syllogism, it is a difficult experimental question to even specify what empirical evidence is required to distinguish between monotonic and nonmonotonic reasoning in the syllogistic fragment. It has been assumed that merely instructing different reasoning criteria is sufficient to discriminate. The empirical problems of discriminating these goals has been largely ignored or denied, and their neglect stems directly from conflict of this difficulty of observation with the descriptivism which we lament. Once a formal specification of an alternative interpretation of the task is available, it is possible to launch a genuine empirical exploration of what participants may be trying to do.

It is not difficult to see why a multiple-logics stance defuses accusations of prescriptive normativism. As soon as there is explicitly acknowledged plurality, then the need for specification of appropriateness conditions for the different logics is clear for all to see. Fortunately, multiplicity brings with it the materials for an answer. Why is classical logic a good model for adversarial reasoning such as the settlement of dispute? Well, it is bivalent, admitting no intermediate truth values. It is extensional, which means the relevant questions of meaning are easily identified, if not necessarily decided, in agreeing premises. It is truth functional, with similar consequences—no hidden meanings can obscure the connection intended by an intensional conditional. It reasons from identified premises with fixed interpretations. Wandering premises are not good for dispute resolution. But above all, its concept of validity requires the preservation of truth in conclusions from true premises under *all* assignments of truth values.

Why is Logic Programming a good logic for cooperative reasoning about the effect on our preferred model of knowledge rich interpretation of new information? Well, the knowledge-base of conditionals corresponds to the long term regularities in the environment, along with the numerous exceptions to these regularities. Working memory holds the representation of the current preferred model of the focal situation (the "closed world"). The closure of the world is made possible by the restriction of expression which allows the rapid settlement of whether a particular proposition can be derived from the large knowledge base. And so on. Even these partial descriptions of the differences between the logics are enough to explain for many contexts whether classical or a nonmonotonic logic is appropriate. The norm can be seen to be appropriate to the goal. It is when human reasoning is assumed to be logically homogeneous, lack of adequate justification is inevitable. For example, there is a pervasive though not universal view in the psychology of reasoning that monotonic and nonmonotonic logics are two ways of "doing the same thing," where the nonmonotonic logic is seen as a poor man's approximation to classical logic. For example, Mental Models theory correctly asserts that to achieve classical reasoning, participants should consider all models of the premises in syllogistic reasoning. But when it is clear that they mostly actually only consider one model, this is considered a performance error (forgetfulness): not a symptom of nonmonotonic goals to identify a preferred model. This is accompanied by separate experimental demonstrations that participants *can* successfully search for counterexample models when explicitly instructed to do so, in a quite different task. This is taken as supporting that indeed the failure to look for them in solving syllogisms is a performance error. At no point is it questioned whether the participants' goal is different in these two tasks. Just because people can do

counterexample reasoning sometimes, does not mean that this is always their goal.

The LP machinery may often operate below awareness; this does not mean that the participant who adopted the goal that it performs does not "have" the goals under which it operates. And plurality is absolutely required for other reasons. There is no way that any logic can provide a model of both dispute and exposition because the logical properties listed above are incompatible[6].

From these arguments it follows that pure descriptivism is impossible in situations where both CL and LP are live options for participants' interpretation (most laboratory reasoning tasks) because choice of logic, and with it reasoning goals, is required for interpretation of the data. There is no alternative to seeking evidence for which goals the participant has adopted (usually inexplicitly). Merely varying the instructions is not an adequate tool for discovery.

3.2. DESCRIPTIVIST APPROACHES TO THE SYLLOGISM CANNOT DISCRIMINATE THESE GOALS

There are 64 pairs of syllogistic premises which can be enumerated with their valid conclusions. There are a some logical glitches about exactly what ought to be listed as valid [7]. The conventional task for studying "syllogistic reasoning" is defined by the goal of "getting these answers" to the question "What follows from these premises?" For example, if the premises are *All A are B. All B are C* then *All A are C* is a valid conclusion. So participants who answer with this conclusion score a point. This is OK as far as it goes as an denationalization, but if it is all we can offer, then it makes the syllogism an uninteresting pursuit for the researcher and participant alike. Who says these ones are valid? So it is generally further assumed by the experimenter that these right answers are given by classical logic—was not Aristotle, the author of the first logical theory of syllogisms, thereby the inventor of classical logic?—but pure descriptivism is already out the window. CL has constitutive norms, and with them its users and uses acquire regulative norms.

Troubles compound. These participants have been selected for not knowing explicitly what the syllogism, or classical logic, are. It is true that they know the natural language of the premises, and it is easy to suppose that this determines the reasoning goal. But it is the *discourse* that they have trouble understanding out of context. And they often complain about the bizarreness of the discourse in ways that make one think they in fact adopt a goal quite different to the one the experimenter stipulates. For example, given *Some A are B. Some C are B* they frequently complain that "it doesn't tell me whether the Bs are the same or different." This complaint makes no sense if the premises are understood "classically." Classically it is absolutely clear that they could be either the same or different unless the quantifiers force them to be related, and in this case they "obviously" do not. Yet about 60% of participants claim that there is a valid conclusion here[8] On a "story-understanding" LP interpretation, they are of course right that the discourse is "defective" and there are ways of fixing it so that there are valid conclusions based on preferred models—several ways.

So we do not yet know what the participants' goals are at any level beyond assuming they are to please the experimenter, who has not been good enough to divulge his goals in a way that the participant can interpret them. Just saying "I want what logically follows" or "what must be true" is not helpful, since "logically" has many meanings in the vernacular ("reason carefully" is often a good gloss), and any participants who have taken intro logic have been weeded out. "Logically" also has many technical meanings. In LP, a conclusion *must* be true (in the current context) if it follows in the current context from the preferred model. The psychological effects of this kind of emphatic instruction are congruent with the idea that participants take a little more care with whatever goals they happen to have.

Why should we care? What clarification of the goals of the participants would make the syllogism more interesting? We should care about the syllogism because it is a suitable microcosm for seeking the psychological foundations of classical logical reasoning, if any, and that is interesting because classical logic is a crucial mathematical model of dispute or demonstration. So we should be interested in how we can characterize reasoning in this task in a way that it will bear some useful relation to reasoning outside this tiny domain, in say first-order classical logic, or even the much smaller, monadic first-order logic. This would be interesting. Tasks are not themselves interesting if there is no way of connecting them outside the laboratory or across domains. Small fragments are good for satisfying the exigencies of experiment, but they are of little interest in themselves. A good fragment generalizes—and for that one needs to know the goals (and norms) of the participant. There are also significant practical educational gains in understanding exactly why it is that participants have trouble differentiating the discourses of two logics. These problems are close to well known problems of mathematics education in distinguishing generation of examples from that of proofs (Stenning, 2002, chapter 5).

The real problem in this example is that there is more than one systematic reasoning goal that participants might adopt in doing the task as set—that is, more than one logic that might apply. The complaint quoted above is one clue here, though there are many others. The complaint is consistent with the idea that participants are adopting what might be called a "story understanding" task: roughly "What is the model of these premises which their author (presumably the experimenter) intends me to understand by them?" In nonmonotonic logics that capture this reasoning process, these are usually referred to as the *preferred* model (Shoham, 1987). This is cooperative nonmonotonic reasoning *to* a unique minimal model (i.e., *one* interpretation of the premises), as opposed to the adversarial monotonic reasoning from an interpretation, to conclusions true in *all possible models*, that classical logic specifies.

[6]Logicians produce "embedding theorems" which prove that one logic can be "embedded" within another, often when the two look rather incompatible. It does not follow that the more encompassing logic is an appropriate cognitive model for the encompassed systems' cognitive applications.

[7]These "glitches" turn out to be at the heart of some of the psychological issues about CL: more below.

[8]Percentage responses here and following are taken from the metanalysis by Khemlani and Johnson-Laird (2012).

The proposal that cooperative communication worked through the contruction by speaker and hearer of what is now known as a "preferred model" appeared in Stenning (1975) and was condensed in Stenning (1978). Nonmontonic logic was new (McCarthy, 1980), and preferred models had to wait several more years (Shoham, 1987), but what was proposed informally was a direct route to cooperation for psychological process accounts (rather than an indirect Gricean pragmatics founded on adversarial classical logic). Stenning and Yule (1997) showed how subtle is the empirical discrimination of reasoning in classical logic and reasoning in nonmonotonic logic in the microcosms of the syllogism. The "Source-Founding Model" described there is a "shell" for capturing syllogistic reasoning processes, and it demonstrated that adopting a "guess the intended model" reasoning goal could actually yield all and only valid classical logical conclusions if the right model (roughly the "weakest") was chosen, without any conceptual change to a new logic. The interesting psychological conceptual problems are about bald conceptual differences, but are actually difficult to resolve experimentally because the syllogism is so inexpressive. There is considerable evidence that most of the success participants achieve in syllogistic reasoning is achieved by preferred model construction. This is an example of the central importance of the empirical study of goals to the psychology of reasoning. Evans (2002) picks up the point about monotonic and nonmonotonic goals and about interpretation, but suggests no empirical approach other than variation in narrow instructions (rather than tasks) which Stenning and Yule (1997) showed to be inadequate.

It is an immediate consequence that merely observing scores on the 64 syllogisms under different instructions in the conventional draw-a-conclusion task, will not tell us what logic a participant is reasoning with. We have to address the logical concepts that they have (for example, attitudes to conditionals with empty antecedents—more presently) and with them their processes of reasoning. We beg the reader's patience with some details which are important for understanding the role distinct goals (embodying distinct norms) play. We will use the diagrammatic methods this reference uses, though it also supplies analogous sentential ones. So for example, the syllogism *All A are B. Some C are not B* is represented by **Figure 1**.

In the final diagram, the single cross marks an element which is C but not A or B, which must exist in any model where the premises are true[9]. The choice of preferred models in the diagrams of each premise, combines with this construction of all consistent sub-regions, and with the rules for retaining or deleting the crosses, to ensure the result that any remaining cross represents an arbitrary individual with the properties defined by its subregion. The surprise is that this individual classically must exist if the premises are true. That is, the rules for choosing the nonmonotonically "preferred" model can conspire, in this tiny fragment of classical logic, to choose a model for the premises

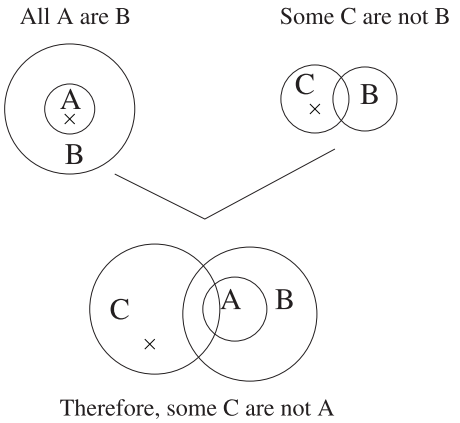


FIGURE 1 | Two premise diagrams unified in the Euler's Circles system of Stenning and Yule (1997). The crosses mark non-empty subregions. In the unified diagram, the A and C circles must be arranged to create the maximum number of minimal sub-regions compatible with the premises. In this case the A and C circles must intersect. Crosses whose minimal sub-region in the premise diagram have been bisected in this unification operation are deleted. Remaining crosses mark minimal models, and thereby indicate classically valid conclusions.

which has to exist in any situation where premises are true i.e., is a classically valid conclusion. This is of course not to say that participants who adopt a generally nonmonotonic goal for the task will automatically adopt the particular procedures required for getting classically valid preferred models: there are many parameterizations of the tweaking of nonmonotonic strategy. Informally, participants have to prefer the "weakest" model.

Stenning and Yule (1997) also provides a sentential algorithm which mirrors this graphical algorithm, as well as a "Source-Founding method" which is an abstract algorithm which captures what is in common between nonmontonic and classical methods. It shows the equivalence of the model manipulations in the diagrams with Aristotle's *ekthesis*. So it will be impossible to empirically distinguish participants' with classical norms from those with these "correctly tweaked" nonmonotonic reasoning norms by merely inspecting input premises and output conclusions. Yet identifying these norms is just what we argued psychology has to do to establish what implicit grasp of classical logic its participants have.

But help lies at hand. What has happened, in our nonmonotonic alternative method, to all those paradoxical properties of classical logic that bother every introductory logic student so much? For example, the paradoxes of material implication, whereby, from $\neg p$ it follows that $p \rightarrow q$; and from q it also follows that $p \rightarrow q$. Or, for a related example, the conclusion that the King of France has been bald since the Revolution because there has been no King of France?: the problem of existential presuppositions. Besides, if the nonmonotonic tweaks get the classical answers, who needs to put up with these crises of classical logic?

So what is the psychological bottom line? The psychological half-way line, is that who needs classical logic is anyone who wants to go beyond the syllogism into the vastly more expressive first-order logic, and needs this still important model of

[9]The diagrammatic system is described in more detail in the reference above and also in Stenning and Oberlander (1995), e.g., Figure 2. In the variant used here, existential presuppositions are made for universals, because that assumption is commonplace in the psychology literature. Below we see that it is not clearly the right assumption when the task context changes to dispute.

demonstration and rational dispute (e.g., for mathematics, science, politics or the law). An experimenter might be tempted to the conclusion that this was just a bad fragment to pick, and progress to the psychological study of first-order or at least monadic first-order logic. There are formidable obstacles on that path, and no one has ventured down it far. But there is an alternative strategy within the syllogism. How can we get data richer than input-output pairings of premise-pairs and conclusions? If the conventional psychological task of presenting a pair of premises and asking whether any, and which of, the eight conclusions follows, brings forth nonmonotonic norms (albeit sometimes refined ones) from most participants, then perhaps what is needed is a new task and task context (dispute perhaps)? And what about getting participants to perform not just inferences, but also *demonstrations* of those inferences (by producing counterexamples)? This would provide evidence beyond input-output functions.

What are the quintessential features of classical reasoning that we should focus on in the data? The clues are in the paradoxes, though it requires some digging to unearth them. We are claiming, as is commonplace in traditional logical discussion, that classical logic is a model of dispute. What does this mean? Its concept of validity is that valid conclusions must be true in all models of the premises. What this means is that there must be no counterexamples (or "countermodels"). So classical logical demonstration is a doubly negative affair. One has to search for the *absence* of *counter*examples, and what is more, search exhaustively. A dispute starts from agreed and fixed premises, considers all situations in which these are all true, and wants to be certain that inference introduces no falsehood. The paradoxes of material implication immediately disappear. If p is false, then $p \supset q$ cannot be false (its truth-table reveals that it can only be false if both p is true and q is false. (And truth tables is all there is to truth-functions). And the same if q is true. So given that p is false or q is true, we cannot introduce falsehood to true premises by concluding q from $p \supset q$. Everything follows from the nature of this kind of dispute, in which the premises must be isolated from other knowledge because they must be explicitly agreed, and in which no shifting of interpretation can be hidden in implications, or indeed in predicates. This latter is ensured by extensional and truth-functional interpretation. The "paradoxes" are thus seen as paradoxical only from the vantage point of nonmonotonic reasoning (our usual vantage point), whose norms of informativeness they violate. In dispute, proof and demonstration, the last thing one wants is the informativeness of new information smuggled in. And if you are engaged in telling a story, *failing* to introduce new information in each addition to the story will invoke incomprehension in your audience. Tautologies do little for the plot. This contrast is what we mean by each logic having its own discourse, and these two are incompatible.

Bucciarelli and Johnson-Laird (1999) earlier presented counterexample construction as an explicitly instructed task using syllogisms, though with a different partly graphical presentation of situations. Their purposes were to refute the claims of Polk and Newell (1995) that in the conventional draw-a-conclusion task, participants do not search for counterexamples, as mental models theory claimed that they understood that they *should*: 'If people are unable to refute conclusions in this way, then Polk and Newell (1995) are certainly correct in arguing that refutations play little or no role in syllogistic reasoning' (Bucciarelli and Johnson-Laird, 1999, page 270). Whilst their investigations of explicit countermodeling do, like ours, establish that participants can, when instructed, find countermodels above chance, they certainly do not counter Polk and Newell's claim that participants do not routinely do this in the conventional task on which mental models theory is based. Other evidence for Polk and Newell's skepticism now abounds (e.g., Newstead et al., 1999). But nowhere do any of these authors explicitly consider whether the participants' *goals* of reasoning in countermovement diverge from their *goals* of reasoning in the conventional task, even less whether they exemplify two different logics. At this stage, Mental Models theory was seen by its practitioners as the "fundamental human reasoning mechanism." Another example of our dictum that it is exactly where homogeneity of reasoning is proposed, that normativism goes off the rails.

Searching for an absence of counterexamples then, is the primitive model-theoretic method of proof in the syllogism classically interpreted. The whole notion of a counterexample to be most natural, and best distinguished from an exception, needs a context of dispute. How do we stage one of those in the lab? Well, we tried the following (Achourioti and Stenning, in preparation). A nefarious character called Harry-the-Snake is at the fairground offering bets on syllogistic conclusions. You always have the choice of refusing the bets Harry offers, but if you think the conclusion he proposes does not follow from his premises (i.e., is invalid), then you should choose to bet against him. If you do so choose, then you must also construct a counterexample to his conclusion. Evidently we also have to explain to participants what we mean by a counterexample (a situation which makes both premises true and the conclusion false); what we mean by a situation (some entities specified as with or without each of the three properties A, B and C; and how to construct and record a counterexample. (In fact we use contentful material that does not affect likelihoods of truth of premises). Two features of this situation are that Harry-the-Snake is absolutely not to be trusted, and that it is adversarial—he is trying to empty your wallet. Another is that you, the participant, have chosen to dispute the claim Harry has made. You do not have to ask yourself "What if I thought this did not follow?" It has a vividness and a directness which may be important. Our selection of 32 syllogisms (unlike Bucciarelli and Johnson-Laird's) was designed to concentrate on the "no valid conclusion" problems which are at the core of understanding CL, and to allow analysis of the "mismatching" of positive and negative middle terms.

Our most general prediction was an increased accuracy at detecting non-valid conclusions. In the conventional task this is extremely low (37%): highly significantly worse than chance: in the new task it is 74%, significantly better than chance, and valid problems are 66% correct, which is also above chance. Valid problems are now harder, but the task now focusses the participant on the task intended. We also made some more specific predictions about a particular class of syllogisms which we call "mismatched," in which the B-term is positive in one premise and negative (i.e., predicate-negated) in the other. Mismatching middle-term

double-existential problems (e.g., *Some B are A, Some C are not-B*) "obviously" do not have single-element models, and so no valid conclusions. Compare a corresponding matched case *Some B are A, Some C are B* which yields as a unification model the single-element: (ABC). The most popular conclusion is *Some C are A*, drawn by 39% of participants. Note that this unification model is *not* a countermodel of this conclusion. With the mismatched example above, one cannot get a 1-element model. This difference between matched and mismatched double-existential problems and their most popular conclusions is systematic, as we describe below.

One might suppose that absence of valid conclusions is a general property of mismatching syllogisms because of the unification barrier to 1-element models, until one thinks about what happens if the first premise was instead *All B are A*. This universal premise would be satisfied by a single element model (such as *A not-B C*). But only if the negated B term is accepted as making the universal premise true by making its antecedent empty. That is, by the very same model which countermodels the existential case. Here is one place where the connection between CL's "paradoxes" and matching/mismatching shows up. Participants accepting the empty antecedent conditional as true can produce this one-element model.

So mismatching may serve as a tracer for issues with empty-antecedents. To find 1-element models for these mismatched problems requires accepting empty-antecedent conditionals as true. Now comes the question, do any of these syllogisms have valid conclusions? They can have 1-element models if one accepts empty antecedent conditionals, but are these models ones that establish valid conclusions? This model does not establish a valid conclusion anymore than the model (ABC) establishes a conclusion for *Some A are B. Some B are C*. In fact the problem does have a different valid conclusion *Some A are not C*. In summary, these mismatched problems provide a way to gain information about participants' intuitive grasp of empty-antecedent conditionals. And accepting empty-antecedent conditionals as true is a special case of accepting the paradoxes of material implication—the essential example of CL's "weirdness"—in the context of dispute. This is what we mean by looking for its "weirdnesses" as being the best evidence of implicit grasp of a logic. CL is weird in disputes; only from the non monotonic perspective, even for "logically naive" subjects.

If a participant has some implicit grasp of the one-element model generalization, and is happy with models satisfying conditionals by making their antecedent empty, then mismatched problems could behave differently than matched in this model-theoretic search-for-counterexample method: the striking logical feature (empty-antecedent conditionals being true) connects directly to an unexplored psychological feature. Mismatched problems, when we do the analysis, are actually observed to be slightly but significantly *harder* than matched ones in the conventional task of constructing a conclusion. To see how they might behave differently in countermodel search, one also needs to consider what the favorite conclusions are in the conventional task. For our example, the favorite response is *No C are A*. Now, we observe, that the model one gets by unifying the premises is (A not-B C) is immediately a *countermodel* of this popular conclusion (ie. some C *are* A in this model). If we take the matched and the mismatched problems in our experimental sample of 32, each paired with its favorite conclusion (from the meta-analysis), we find all the mismatched problems have this property that the unification model countermodels the favorite (and usually invalid) conclusions; whereas with the matched problems, the unification model is, in each case a *model* of the erroneous but favorite conclusion. This is evidently an empirical psychological generalization (favorite conclusions in a particular task have no logical status), though we clearly need the CL model-theory to even notice this piece of psychology. We predicted that *when looking for countermodels* (ie. doing CL), mismatched problems should be easier than mismatched ones.

What actually happens when Harry shows up? To cut a long story short, participants experience disputing with Harry-the-Snake as a much more arduous task than the conventional draw-a-conclusion task. They slow down by a factor of about three, an observation that already casts doubt on claims that this countermodel search takes place in the conventional task. Countermodel reasoning is hard work. Their overall accuracy of judgment of validity is not hugely increased, but it does not suffer from the extreme asymmetry of the conventional task. Both VC and NVC problems are done at a better than chance level. The control group in our conventional task control group are also much better than the literature average (these are highly selected students), but they are still asymmetrical in their success in the same way with VC easier than NVC problems. So we find the predicted improvement in detecting invalid conclusions, and we find that indeed whereas mismatched problems are somewhat harder than matched ones on the conventional task, they are substantially *easier* in countermodel reasoning in dispute with Harry, and that participants show evidence of accepting empty antecedent conditionals as true in the dispute task.

The pattern of errors in countermodel construction is consistent with a process by which participants first try to construct a premise model, then check to see if it is a countermodel, and if it is not, then adjust it to try to achieve a falsification of the conclusion. The problem appears to be that the adjustment often yields a model that falsifies the conclusion but is no longer a model of the premises. Mismatched models are more accurately countermodeled, and this is because the models that result from the unification of their premises are already countermodels of Harry's proposed conclusions, as illustrated above. This pattern that mismatched problems are actually easier for countermodel construction whereas they are harder in the conventional task strongly suggests that the majority of participants in the conventional task are operating proof-theoretically, probably by the nonmonotonic methods discussed above.

The countermodel construction data provides rich evidence that empty antecedent conditionals can be treated as true in this context. If the data is scored requiring existential presuppositions, most of the models produced for problems with one positive and one negative universal (i.e., no explicit existential premise) are not even models of the premises, let alone countermodels of the conclusion. A final observation that supports this general interpretation of a change of process invoked by dispute with Harry is that the orders of difficulty of problems in the conventional and

in the Harry tasks are actually uncorrelated—an extremely strong result in support of the claim that here is the first task in the literature that produces substantial classical reasoning conducted on a classical conceptual basis. But even here, there are still many errors in countermodel reasoning. The usual justification of the conventional task is that the order of the difficulty of problems is systematic and always the same. The first time anyone makes a comparison with a context designed to invoke a different logic, one finds this order of difficulty changes radically.

Clarifying the intended goals of reasoning (norms to adopt) for participants is one of the few ways we have of pursuing the question whether there are contexts in which participants intuitively understand the concepts of a logic. One can imagine the objection that we have told them to do countermodel reasoning and so it is not surprising that they appear to reason classically. But this is a psychologically bizarre idea. It's no use telling these participants to reason in classical logic because they do not explicitly know what that means. They do have some grasp of what a dispute is, and the role of counterexamples therein—the discourse of dispute. We are merely negotiating a common reasoning norm with our participants. If they did not understand these things, the negotiation would not succeed. We doubt it succeeds with all our participants. But we certainly do not instruct them about what to do with empty antecedent conditionals. And sure enough, we see the peculiarities of classical logical reasoning in their performance. This is just what the psychological foundations of classical logic are: an inexplicit intuitive grasp of dispute. These empirical conceptual questions such as "What do participants 'know' about classical logic?" have far more psychological reach than questions about how many syllogisms do participants get "right" in any particular contextualized task where the goals are not understood the same way by participant and experimenter, or across participants.

Participants are, unsurprisingly, not tactically expert. But here at least is the beginning of an empirical program to study this kind of reasoning in contradistinction to various kinds of nonmonotonic reasoning. Although the two may overlap within the syllogism, outside the syllogism they diverge. And even within the syllogism, here is evidence that the two very different reasoning goals are operative in different contexts, and lead to radically different mental processes, each incomprehensible without an understanding of the different logical goals, and of the participants' informal contextual understandings of their logical goals.

4. REASONERS' GOALS IN THE NEW PROBABILISTIC PARADIGM

Classical logic has been found wanting as a complete model of human inference for many reasons, some of which we have already covered. The "new paradigm" of subjective probabilities aspires to become its replacement (Over, 2009; Oaksford and Chater, 2013). A central question has been whether people's interpretation of indicative conditionals, 'if A, then B', is given by the material conditional $A \supset B$ (see **Table 1** for a reminder of its truth values) or the conditional probability $P(B|A)$. There is evidence that in some circumstances participants do indeed reason that the probability of 'if A, then B' is given by $P(B|A)$, both when dependencies between antecedent and consequent are expressed in the task through joint frequencies about patterned cards (Evans et al., 2003; Oberauer and Wilhelm, 2003) and when dependencies are derived from causal beliefs (Over et al., 2007). These interpretations also extend to conditional bets such as "I bet you 1 Euro that if the chip is square then it is black" (Politzer et al., 2010), a result which is predicted by foundational work on subjective probability by Bruno de Finetti (Milne, 1997, gives an overview).

Table 1 | Truth values of the classical logic material conditional ($A \supset B$), conjunction ($A \wedge B$), and semantic values of the conditional event ($B|A$) and biconditional event ($B|A$) $\wedge$ ($A|B$), where 1 denotes "true," 0 denotes "false," and u denotes "undefined."

A	B	$A \supset B$	$A \wedge B$	$B\|A$	$(B\|A) \wedge (A\|B)$
1	1	1	1	1	1
1	0	0	0	0	0
0	1	1	0	u	0
0	0	1	0	u	u

The conditional event, $B|A$, is often defined only for conditional probabilities in terms of the ratio formula,

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

under the condition that $P(A) > 0$. Coherence-based probability logic (CPL), proposed as a competence model for how people reason (Pfeifer and Kleiter, 2009), makes this a primitive, $B|A$, which is "undefined," "void," or "undetermined" when the antecedent is false, matching how participants often interpret the conditional when reasoning under certainty (Johnson-Laird and Tagart, 1969). Although this interpretation is often called the "defective" conditional, there is a long history of justification suggesting that there is nothing defective about it. CPL derives a semantics for conditional probabilities, providing a bridge between certainty and uncertainty. This explains why people who use a "defective" conditional when reasoning about certainty also reason using conditional probabilities for uncertain conditionals (Evans et al., 2007, show this empirical link): it's the same underlying conditional.

Hailperin (1996) provides a further analysis of this conditional event (he calls it the "suppositional") in terms of a more primitive operator in an extension of classical propositional logic, "don't care" logic. We present this in a some detail here as it shows clearly the relationship with classical logic. Let 1 denote "true," 0 denote "false," and u denote "undefined." The ordering on these semantic values is $0 \leq u \leq 1$. This leads to natural *min* and *max* functions for deciding the minimum and maximum of two values which are used to define conjunction and disjunction, respectively. Let $min(x, y) = z$; then z is either the x or y and chosen such that $z \leq x$ and $z \leq y$, i.e., the value is less than or equal to both x and y according to the ordering above. Let $max(x, y) = z$; then again z is one of the x or y and $z \geq x$ and $z \geq y$, i.e., the value is greater than or equal to both x and y. Some examples to illustrate: $max(0, 1) = 1$ and $min(0, 1) = 0$. If x and y are the same value then the answer is that value for both *min* and *max*. When the u value is included then $max(0, u) = u$ (since u is greater than or equal to both 0 and

itself) and $min(0, u) = 0$ (since 0 is less than or equal to both u and itself). Finally $1 - u = u$ (this is used for defining negation). U is a semantic function from formulas to semantic values, i.e., one of 0, 1, or u, as follows:

$$U(\neg A) = 1 - U(A)$$
$$U(A \wedge B) = min(U(A), U(B))$$
$$U(A \vee B) = max(U(A), U(B))$$

If we use only 1s and 0s, this is also the semantics of classical propositional logic. When u is included, then the semantics is equivalent to Kleene's strong 3-valued logic (Kleene, 1952) which turns out to be useful for the semantics for logic programming (Fitting, 1985). Hailperin (1996) introduces an additional "don't care" unary connective, $\triangle$, with a semantic value defined:

$$U(\triangle A) = \begin{cases} 0, & \text{if } U(A) = 0 \\ u, & \text{otherwise} \end{cases}$$

If A is true then $\triangle A$ evaluates to u; otherwise it has the same semantic value as A. This allows $B|A$ to be defined $\triangle\neg A \vee (A \wedge B)$, giving the same semantic values as CPL. Note the similarity with the disjunctive expression of the material conditional, $\neg A \vee B$, which is equivalent to $\neg A \vee (A \wedge B)$. Both have the same semantic value when the antecedent is true, equivalent to a conjunction. The disjunct highlights the difference when the antecedent is false: for the conditional event the conditional is undefined but for the material conditional it is true. (This is one of many non-classical truth semantics; Baratgin et al. (2013) provide other interesting examples of further logical components which are useful for psychological theorizing.) Individuals and quantifiers are missing from this semantics, which limits its ability to model discourse; for instance it is not clear how to model an interpretation of "most logicians who develop a logic love it."

Returning to the psychology, there are interesting twists to the new paradigm story. It turns out that the experimental data also require us to model a defective biconditional, what Fugard et al. (2011b) named the biconditional event. This is expressed as $(B|A) \wedge (A|B)$ (see **Table 1** for its semantics values) and is equivalent to $A \wedge B|A \vee B$. Developmental studies show that 12 year olds respond mostly with conjunctions, then by age 16 biconditional event interpretations appear before disappearing again in adults (Gauffroy and Barrouillet, 2009). In adults, it is well replicated that nearly half of participants interpret the conditional as a conjunction, $A \wedge B$. Shifts of interpretation have also been found within adults: many participants who begin with a conjunction interpretation change that interpretation (without feedback) to a conditional probability (Fugard et al., 2011b; Pfeifer, 2013). Participants occasionally are explicit about this, describing their reasoning about what they think they are supposed to do and changing their goals, occasionally swearing as they do so, a sure sign of norms awry.

Gauffroy and Barrouillet (2009) explain the developmental trend in a revision of mental models theory. Essentially the idea is that more slots of memory are required as one moves from conjunction—produced by heuristic processes immune to strong developmental changes' (p. 274)—through biconditional event, to conditional event. All reasoners are assumed to have the same reasoning goals, they just fail if they have insufficient memory. Fugard et al. (2011b) instead argued that there are two main stages to reasoning about these sorts of conditionals when the dependencies are expressed in the stimulus, for instance as colored cards. First one has to visually perceive the dependencies, which requires attending to all cases. If you are reasoning about new evidence then you first have to examine the evidence. All evidence is initially relevant, even those cases where the antecedent is false, as you can only tell it is false once you have seen it. The developmental trend can be seen as strategic ignorance when all the evidence has been examined: first from no narrowing of hypothetical scope for conjunctions ($A \wedge B$), to focusing on only those cases where either antecedent or consequent are true ($A \wedge B|A \vee B$), finally to only those cases where the consequent is true, ($A \wedge B|A$) which is equivalent to the conditional event $B|A$. Further support for this model is that conjunctions seem to disappear in Experiment 1 by Over et al. (2007) where instead of reading dependencies from the stimulus, they were taken from beliefs, e.g., that "If nurses" salaries are improved then the recruitment of nurses will increase. There is no need to consider evidence when you are asked your opinion. This hypothetical narrowing could be for many reasons. Perhaps there are variations in pragmatic language function which affect the interpretation of what the experimenter wants. Another explanation is that working memory and reasoning processes have competing goals: represent everything that one sees versus reason about top-down goals concerning the present task (Gray et al., 2003). The two could well be related and influence reasoning about goals. People can switch goals for resource reasons.

The "new paradigm" is often presented as providing the semantics for the conditional as illustrated by 'the Equation': P('if A, then B') $= P(B|A)$. But interpretation is required for probabilities too. Fugard et al. (2011a) showed that a relevance pragmatic language effect, well replicated for non-probability problems in the classical logic paradigm, also affects probabilistic theories of conditionals. Consider the following sentence about a card.

> If the card shows a 2, then the card shows a 2 or a 4.

In the old binary paradigm, people tend to think this sentence is false (though with the usual individual differences) since the possibility that the card could be a 4 seems irrelevant if you know it is a 2. Fugard et al. (2011a) found that when participants were shown four cards, numbered 1 to 4, and told that one has been chosen at random, many thought the probability of this sentence is 0. Probability logic (with the simple substitution interpretation) predicts that they would say the probability is 1. Given the same cards but instead the sentence

> If the card shows a 2, then the card shows an even number,

most participants give the probability 1 which is now consistent with the Equation. The new paradigm of transforming 'if's into conditional events does not predict this different in interpretation. Here, as for much of the psychology of reasoning, there are

differences between participants in interpretation and not all reasoners have the goal to take relevance into consideration. Fugard et al. (2011a) found no association between irrelevance aversion and tendency to reason to a conjunction probability, suggesting that the two processes are logically and psychologically distinct.

The problem for the probability story, as the semantics above shows, is that the disjunction in probability logic is the same as the disjunction in classical logic, so this provides a clue for a solution. Schurz (1991) provided an extension of classical logic for interpretations like these: sentence φ is a *relevant* conclusion from premises Γ if (a) it follows according to classical logic, i.e., $\Gamma \vdash \varphi$ holds, and (b) it is possible to replace any of the predicates in φ with another such that φ no longer follows. Otherwise φ is an irrelevant conclusion. Take for instance the inference $x = 2 \vdash x = 2 \vee x = 4$. Since $x = 4$ can be replaced with any other predicate (e.g., for the synesthetes $red(x)$) without affecting validity, the conclusion is irrelevant. However for the inference $x = 2 \vdash even(x)$, not all replacements preserve validity, for instance $odd(x)$ would not, so the conclusion is relevant. Fugard et al. (2011a) propose adding this to the probability semantics.

Reasoners still have goals when they are reasoning about uncertain information. There are competing processes related to working memory and planning, which could explain developmental processes and shifts of interpretation within participants. Goals related to pragmatic language, such as relevance, are also involved in uncertain reasoning. The investigations above highlight the importance of a rich lattice of related logical frameworks. The problems of classical logic have not gone away since, as we have shown, much of classical logic remains in the 3-valued semantics. Rather than only examining whether or not support is found for the probability thesis, instead different norms are needed through which to view the data and explain individual differences. These norms need to bridge back to the overarching goals reasoners have.

We finish this section with a comment on the treatment of this same problem by Bayesian modeling. The probability heuristic model (PHM) of Chater and Oaksford (1999) was one of the first to protest against the idea that classical logic provided the only interpretation of syllogistic performance. A protest with which we evidently agree. This Bayesian model certainly changes the measures of participants accuracy in the task. For the present argument, two observations are relevant. Firstly, PHM is probably best interpreted as a probability-based heuristic *theorem prover* for classical logic. The underlying logic is still in classical logic and even includes first-order logic statements. The truth of the propositions is assessed classically. This means that despite the rejection of the formal model of classical logic, it has not departed very far. PHM does not propose an alternative interpretation of the goal of reasoning as we do here. Secondly, once the Bayesian model is in place, the psychology stops. There is no motivation for seeking other models of other qualitatively different kinds of reasoning, because probability based models are supposed to account for all reasoning. This may be a consequence of at least a whiff of poor prescriptivism here, and bears out the claim we made that this problem is found wherever one framework is seen as sufficient. In contrast, in a multiple-logics approach, contrast between logics is a rich source of insight and guidance as to how to find the relevant psychological evidence. It should be evident from this example that logic can make empirical experimental analysis much richer. Instead of hundreds of experiments on essentially the same design, one gets a vista of empirical questions to explore.

5. CONCLUSIONS

A variety of formal systems, with their different constitutive norms, and their different consequences for the regulative norms of their users, will be required for modeling the different goals of human reasoning. The main goal of the experimental program of psychology of reasoning and decision at this point should be to find contexts in which participants will exhibit their maximum grasp of each system. Exploration can then spread out to investigate how the logics work together in more complex tasks; how participants can generalist from these focal points; and how teaching affects what they can do. If we win our bet on Harry as a good teacher of an explicit grasp of the logical differences between disputes and stories, and we can show the rudiments of classical logic in a good proportion of participants' performances, then that does not mean that CL "won" over nonmonotonic logics such as LP, or over probability logics, or whatever other logics can be shown to have their contexts. It means we know a little more about where to look for classical logic's psychological roots. We can ask how do these cognitive foundations develop, and what individual and social experiences affect them. We can ask how people at different stages of development and education experience the phenomenology of their reasoning. We can ask how best to achieve educational goals of making explicit students' knowledge of logics. And so on.

In many cases, the empirical discriminations between logics are surprisingly hard. Natural languages often do not provide adequate (or indeed any) cues to intended reasoning goals. People are good at recognizing the goals in customary rich social contexts (few mistake a dispute for a story), but the lab removes all these cues, as do many real-world professional contexts. Much effort is currently going onto the issue of what probability theory is good for, but little into where nonmonotonic logics are to be preferred. Deep knowledge of the logical and computational properties of these systems is available outside psychology but often shunned. Formal systems such as logics and probability are still conventionally seen as competing with psychology for explanations of reasoning. A recent prominent example of this attitude (here to probability rather than logic) is Jones and Love (2011).

> Bayesian modeling of cognition has undergone a recent rise in prominence, due largely to mathematical advances in specifying and deriving predictions from complex probabilistic models. Much of this research aims to demonstrate that cognitive behavior can be explained from rational principles alone, without recourse to psychological or neurological processes and representations.

Bayesians would dispute whether they claim to explain in rational terms *alone*. We would disagree with many of their "rational explanations." One might certainly feel disappointed if rational explanations were all of psychology. One of the reasons for our detailed examples is to show that logical bases for explanations

do not mean they cannot reveal psychological processes. A huge amount of research in a descriptivist style has failed to make the most important empirical distinctions about which interpretations of the tasks are adopted. But having said all this, to challenge the idea that rational explanations are part of psychology is truly extraordinary. What is needed is more attention to norms, and to the way the constitutive norms of formal systems give rise to regulative norms for their use, and above all, on participants' access to these norms of both kinds.

There is no alternative to a psychology of reasoning which has a rich theoretical vocabulary of reasoning norms, which constitute different goals, and a fine nose for finding the contexts of reasoning that call for the goals, based on the norms of the logical models. Descriptivism never worked in any science.

ACKNOWLEDGMENTS

The authors would like to thank the editors, Shira Elqayam and David Over, and Laura Martignon for their insightful comments which we feel have greatly improved the paper.

REFERENCES

Achourioti, T., Fugard, A., and Stenning, K. (2011). Throwing the normative baby out with the prescriptivist bathwater: commentary on Elqayam and Evans. *Behav. Brain Sci.* 34:249. doi: 10.1017/S0140525X1100046X

Anderson, J. (1983). *The Architecture of Cognition*. Harvard University Press.

Baratgin, J., Over, D. E., and Politzer, G. (2013). Uncertainty and the de Finetti tables. *Think. Reason.* 19, 308–328. doi: 10.1080/13546783.2013.809018

Bucciarelli, M., and Johnson-Laird, P. (1999). Strategies in syllogistic reasoning. *Cogn. Sci.* 23, 247–303. doi: 10.1207/s15516709cog2303_1

Chater, N., and Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cogn. Psychol.* 38, 191–258. doi: 10.1006/cogp.1998.0696

Cummins, D. (1995). Naive theories and causal deduction. *Mem. Cogn.* 23, 646–658. doi: 10.3758/BF03197265

Elqayam, S. (2012). Grounded rationality: descriptivism in epistemic context. *Synthese* 189, 39–49. doi: 10.1007/s11229-012-0153-4

Elqayam, S., and Evans, S. (2011). Subtracting ought from is: descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Evans, J., and Over, D. (1996). *Rationality and Reasoning*. Hove: Psychology Press.

Evans, J. St. B. T. (1993). "Bias and rationality," in *Rationality: Psychological and Philosophical Perspectives*, eds K. I. Manktelow and D. E. Over (London; New York: Routledge), 6–30.

Evans, J. St. B. T. (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychol. Bull.* 128, 978–996. doi: 10.1037/0033-2909.128.6.978

Evans, J. S. B. T., Handley, S. J., and Over, D. E. (2003). Conditionals and conditional probability. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 321–335. doi: 10.1037/0278-7393.29.2.321

Evans, J. S. T., Handley, S. J., Neilens, H., and Over, D. E. (2007). Thinking about conditionals: a study of individual differences. *Mem. Cogn.* 35, 1772–1784. doi: 10.3758/BF03193509

Fitting, M. (1985). A Kripke-Kleene semantics for logic programs. *J. Logic Prog.* 2, 295–312. doi: 10.1016/S0743-1066(85)80005-4

Fugard, A. J. B., Pfeifer, N., and Mayerhofer, B. (2011a). Probabilistic theories of reasoning need pragmatics too: modulating relevance in uncertain conditionals. *J. Pragmat.* 43, 2034–2042. doi: 10.1016/j.pragma.2010.12.009

Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011b). How people interpret conditionals: shifts toward the conditional event. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 635–648. doi: 10.1037/a0022329

Gauffroy, C., and Barrouillet, P. (2009). Heuristic and analytic processes in mental models for conditionals: an integrative developmental theory. *Dev. Rev.* 29, 249–282. doi: 10.1016/j.dr.2009.09.002

Gigerenzer, G. (2001). Content-blind norms, no norms, or good norms? a reply to vranas. *Cognition* 81, 93–103. doi: 10.1016/S0010-0277(00)00135-9

Gigerenzer, G., and Selten, R. (2001). "Bounded rationality: the adaptive toolbox," in *Dahlem Konferenzen* (MIT Press).

Gray, J. R., Chabris, C. F., and Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* 6, 316–322. doi: 10.1038/nn1014

Hailperin, T. (1996). *Sentential Probability Logic. Origins, Development, Current Status, and Technical Applications*. Bethlehem: Lehigh University Press.

Hertwig, R., and Todd, P. M. (2005). "More is not always better: the benefits of cognitive limits," in *Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making*, chapter 11, eds D. Hardman and L. Macchi (Wiley), 213–231. doi: 10.1002/047001332X.ch11

Johnson-Laird, P., and Tagart, J. (1969). How implication is understood. *Am. J. Psychol.* 82, 367–373. doi: 10.2307/1420752

Jones, M., and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behav. Brain Sci.* 34, 169–188. doi: 10.1017/S0140525X10003134

Khemlani, S., and Johnson-Laird, P. N. (2012). Theories of the syllogism: a meta-analysis. *Psychol. Bull.* 138, 427–457. doi: 10.1037/a0026841

Kleene, S. C. (1952). *Introduction to Metamathematics*. New York, NY: Van Nostrand.

MacGregor, J. N. (1987). Short-term memory capacity: limitation or optimization? *Psychol. Rev.* 94, 107–108. doi: 10.1037/0033-295X.94.1.107

Martignon, L., and Krauss, S. (2003). "Reconciling bayesianism and bounded rationality," in *Emerging Perspectives on Judgment and Decision Research*, eds S. L. Schneider and J. Shanteau (Cambridge: Cambridge University Press), 108–122. doi: 10.1017/CBO9780511609978.006

McCarthy, J. (1980). Circumscription – a form of non–monotonic reasoning. *Art. Intell.* 13, 27–39. doi: 10.1016/0004-3702(80)90011-9

Mercier, H., and Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–74. doi: 10.1017/S0140525X10000968

Milne, P. (1997). Bruno de Finetti and the logic of conditional events. *Br. J. Philos. Sci.* 48, 195–232. doi: 10.1093/bjps/48.2.195

Newstead, S. E., Handley, S. J., and Buck, E. (1999). Falsifying mental models: testing the predictions of theories of syllogistic reasoning. *Mem. Cogn.* 27, 344–354. doi: 10.3758/BF03211418

Oaksford, M., and Chater, N. (1991). Against logicist cognitive science. *Mind Lang.* 6, 1–38.

Oaksford, M., and Chater, N. (1998a). *Rationality in an Uncertain World: Essays on the Cognitive Science of Human Reasoning*. Hove: Psychology Press/Erlbaum (UK) Taylor & Francis. doi: 10.4324/9780203345955

Oaksford, M., and Chater, N. (1998b). "A revised rational analysis of the selection task: exceptions and sequential sampling," in *Rational Models of Cognition*, eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 372–398.

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford University Press.

Oaksford, M., and Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Think. Reason.* 19, 346–379. doi: 10.1080/13546783.2013.808163

Oberauer, K., and Wilhelm, O. (2003). The meaning(s) of conditionals: conditional probabilities, mental models, and personal utilities. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 680–693. doi: 10.1037/0278-7393.29.4.680

Over, D. E. (2009). New paradigm psychology of reasoning. *Think. Reason.* 15, 431–438. doi: 10.1080/13546780903266188

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., and Sloman, S. A. (2007). The probability of causal conditionals. *Cogn. Psychol.* 54, 62–97. doi: 10.1016/j.cogpsych.2006.05.002

Pfeifer, N. (2013). The new psychology of reasoning: a mental probability logical perspective. *Think. Reason.* 19, 329–345. doi: 10.1080/13546783.2013.838189

Pfeifer, N., and Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *J. Appl. Logic* 7, 206–217. doi: 10.1016/j.jal.2007.11.005

Politzer, G., Over, D. E., and Baratgin, J. (2010). Betting on conditionals. *Think. Reason.* 16, 172–197. doi: 10.1080/13546783.2010.504581

Polk, T. A., and Newell, A. (1995). Deduction as verbal reasoning. *Psychol. Rev.* 102, 533–566. doi: 10.1037/0033-295X.102.3.533

Schurz, G. (1991). Relevant deduction. *Erkenntnis* 3, 391–437.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139173438

Shoham, Y. (1987). "A semantical approach to non-monotonic logics," in *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (Milan), 388–392. Reprinted in Ginsberg (1987).

Simon, H. A. (1972). Theories of bounded rationality. *Decis. Organ.* 1, 161–176.

Stanovich, K. E. (2009). *Decision Making and Rationality in the Modern World (Fundamentals in Cognition)*. New York, NY: Oxford University Press.

Stenning, K. (1975). *Understanding English Articles*. Ph.D. thesis, Rockefeller University.

Stenning, K. (1978). "Anaphora as an approach to pragmatics," in *Linguistic Theory and Psychological Reality*, eds M. Halle, J. Bresnan, and G. Miller (Cambridge, MA: MIT Press).

Stenning, K. (2002). *Seeing Reason. Image and Language in Learning to Think*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198507741.001.0001

Stenning, K. (2012). "Multiple logics within argument: how defeasible and classical reasoning work together," in *Computational Models of Argument: Proceedings of COMMA 2012*, eds B. Verheij, S. Szeider, and S. Woltran (IOS Press), 14–20.

Stenning, K., and Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: logic and implementation. *Cogn. Sci.* 19, 97–140. doi: 10.1207/s15516709cog1901_3

Stenning, K., and van Lambalgen, M. (2001). Semantics as a foundation for psychology. *J. Logic, Lang. Inf.* 10, 273–317. doi: 10.1023/A:1011211207884

Stenning, K., and van Lambalgen, M. (2004). A little logic goes a long way: basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cogn. Sci.* 28, 481–530. doi: 10.1207/s15516709cog2804_1

Stenning, K., and van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT University Press.

Stenning, K., and Yule, P. (1997). Image and language in human reasoning: a syllogistic illustration. *Cogn. Psychol.* 34, 109–159. doi: 10.1006/cogp.1997.0665

Stich, S. (1990). *The Fragmentation of Reason*. MIT Press.

Störring, G. (1908). Experimentelle untersuchungen über einfache schlussprozesse. *Arch. Gesamte Psychol.* 11, 1–127.

Todd, P. M., Gigerenzer, G., and the ABC Research Group. (2012). *Ecological Rationality: Intelligence in the World*. Oxford University Press. doi: 10.1093/acprof:oso/9780195315448.001.0001

Wason, P. C. (1968). Reasoning about a rule. *Q. J. Exp. Psychol.* 20, 273–281. doi: 10.1080/14640746808400161

Wetherick, N. (1970). On the representativeness of some experiments in cognition. *Bull. Br. Psychol. Soc.* 23, 213–214.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 24 September 2014; published online: 20 October 2014.

Citation: Achourioti T, Fugard AJB and Stenning K (2014) The empirical study of norms is just what we are missing. Front. Psychol. ***5****:1159. doi: 10.3389/fpsyg.2014.01159*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Exploring normative models

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 23 May 2014
doi: 10.3389/fpsyg.2014.00488

Development and necessary norms of reasoning

Henry Markovits*

Département de Psychologie, Université du Québec à Montréal, Montréal, QC, Canada

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Jonathan St. B. T. Evans, University of Plymouth, UK
Andy Fugard, University College London, UK

****Correspondence:***
Henry Markovits, Département de Psychologie, Université du Québec à Montréal, C.P. 8888, Succursale "A", Montréal, QC H3C 3P8, Canada
e-mail: henrymarkovits@gmail.com

The question of whether reasoning can, or should, be described by a single normative model is an important one. In the following, I combine epistemological considerations taken from Piaget's notion of genetic epistemology, a hypothesis about the role of reasoning in communication and developmental data to argue that some basic logical principles are in fact highly normative. I argue here that explicit, analytic human reasoning, in contrast to intuitive reasoning, uniformly relies on a form of validity that allows distinguishing between valid and invalid arguments based on the existence of counterexamples to conclusions.

Keywords: development, reasoning, logic, norms, communication, epistemology

The question of the potential usefulness of normative models in understanding human reasoning is a complex one, something that underlies some of the more important debates in the psychology of reasoning. Some of the earliest debates about the nature of human reasoning were explicitly framed around the question of whether human reasoning is essentially "logical" (e.g., Henle, 1962). In these debates, the logical position essentially claimed that humans possessed an inferential apparatus that would (mostly) invariably lead to inferences that corresponded to those found in elementary logic textbooks, reprising Boole's view that Boolean logic simply described human reasoning. A more nuanced approach to this question was given by Braine's (1978) theory that claimed that humans possessed certain limited syntactic reasoning procedures that invariable led to "logically correct" inferences (see also Rips, 1983). These inference rules were the product of biological evolution. Finally, Piaget's theory (Inhelder and Piaget, 1958) made a different claim, suggesting that while children went through stages where their reasoning was constrained by physical and concrete parameters, their development led more or less invariably to the stage of formal reasoning, where logical reasoning is the norm. In fact, Piaget explicitly proposed propositional logic (albeit a modified version of this) as a competence model for formal thought.

Unfortunately for these approaches, empirical research has clearly shown that human inferential performance is highly variable (Markovits, 1985; Overton et al., 1987; Cummins et al., 1991). Many studies have shown that when even educated adults are given what appear to be formally identical arguments, they give difference conclusions. Judgments of deductive validity differ as a function of premise content (e.g., Markovits and Vachon, 1990; Thompson, 1994; Cummins, 1995), and in response to factors such as conclusion believability (Evans et al., 1983). There is little surface evidence that the use of classical propositional logic as a consistent basis for inferential reasoning is very wide-spread, even among highly educated populations. One reaction to these studies has been an attempt to reject the idea that human reasoning is logical at all, by suggesting that much of the inferential apparatus is dominated by biologically based forms of inference. For example, the heuristics described by Tversky and Kahneman (2004) and Gigerenzer and Selten (2001), although differing in many respects provide simple, context-specific forms of rapid inferential reasoning. These heuristics are context dependent, and their use can account for at least some of the variability in human reasoning. However, they do not correspond to a clear model of logic of any kind, although one might suppose (as Gigerenzer explicitly argues) that they are biologically efficient. Similarly, the probabilistic model proposed by Oaksford and Chater (2003, 2007) and Evans et al. (2007) suggests that inferential procedures model the (Bayesian) statistical properties of people's knowledge of their environment. Such models propose that people process relations in a way that explicitly reflects their personal beliefs, which in turn is at least partly determined by real-world knowledge stored in long-term memory (Oaksford and Chater, 2012). Inferences are thus basically probabilistic, and essentially variable, and translate the real nature of people's underlying knowledge. The question of whether reasoning of this kind can be cast in terms of a normative model is open, partly because there is not a strong consensus about the way that probabilistic models function (Elqayam and Evans, 2013; Oaksford and Chater, 2013).

Nonetheless, it is worth making one specific point in this context. Probabilistic models propose that people's inferences are determined by their individual estimations of conclusion likelihood. Since there is no mechanism by which such estimations can be judged as being more or less accurate, a normative model that depends on some external criteria might seem to be impossible to verify. It might, however, be possible to model standard deductive inferences within a Bayesian framework. Deductive reasoning can be seen as an attempt to construct a representation of premises for which there is a shared attempt to maintain some consistent level of internal probability, e.g., a shared belief that the probability of $q|p$ for a given major premise is close to 1 (Oaksford and Chater, 2012). In this case, it might be possible to use a normative model in order to evaluate the way that people reason in this constrained system. However, since such an exercise is

clearly artificial, and does not generally reflect the nature of real world information, norms of this kind will be correspondingly artificial.

The key point is that probabilistic models of inference essentially depend on what must be idiosyncratic representations of real world probabilities, since they depend on information stored in long-term memory. This is quite critical, since it makes Bayesian norms almost by definition undetectable. Bayesian models are used to understand how people can detect environmental regularities, something that is clearly biologically useful, since it allows some level of anticipation of the specific properties of a person's immediate environment (Tenenbaum et al., 2006). However, environments can be variable and individual experience will reflect this variability. Thus, probabilistic models produce by their very nature variable outputs that cannot be compared since this variability reflects variability in inputs. Inferential reasoning can be applied across the whole range of experience and probabilistic approaches to inference must then reflect the wide variety of individual experience. Thus, it could be argued that these approaches suggest that human reasoning cannot, even in principle, be described by a normative model (Elqayam and Evans, 2011).

In the following, I will nonetheless attempt to argue that despite variability and the undoubted influence of many forms of heuristics, human reasoning in its conscious component does indeed depend on a simple normative form of basic logic (which does not necessarily correspond to a specific logical model), for both epistemological and developmental reasons.

WHAT IS A NORMATIVE MODEL?

Before attempting a more specific analysis, it is important to make some initial distinctions. Normative models can be considered in very different perspectives (Elqayam and Evans, 2011). In the following, I will consider that a normative model is a prescriptive description of the *optimal* way that a system should function in order to accomplish its basic goals. It is important to distinguish between such models and descriptive models, which are attempts to describe the actual workings of a given system in a real-life situation. Simple variability is not an indication that a normative model is inapplicable to a given system. However, variability requires showing that the system's functioning tends towards a normative model when conditions are optimized.

There is one further critical part of any analysis of normative models. Normative models are mathematical or logical abstractions that aim to capture the essential functioning of what are necessarily messy and complex systems. Such models are, by definition, the product of human reasoning, since they are the result of people trying to understand the basic parameters of a specific system. The role of normative models in the understanding of human reasoning becomes double-edged, since such models must not only describe the way that people can optimally reason, but importantly these models must also be able to account for the ability of people to construct these models in the first place.

With this in mind, it is useful to note that many normative models have epistemological underpinnings that are essentially based on standard bivalent deductive logic. Given the increasing importance of Bayesian models, as discussed previously, it is particularly useful to note that Bayesian statistics are derived using such logic. In fact, careful analysis of the *arguments* for probabilistic models clearly shows that these are based not on Bayesian inferences, but on classical logical arguments. If human inferences were uniformly Bayesian, then one would expect arguments to be phrased specifically in terms of degree of belief. However, conclusions that explicitly leave open the possibility that alternative theories have a clear probability of being correct are rarely encountered. In other words, it is important to distinguish between the characteristics of the output of a given model, and the epistemological underpinnings of these models. In most fields, the second part of this equation is basically irrelevant. When discussing characteristics of normative models of human reasoning, this becomes fundamental. In fact, one key component of the argument that will be presented is that a minimal normative model for human reasoning is necessary in order to account for the ability to produce normative models in the first place.

DUAL-PROCESS THEORIES

There is one final distinction that is important in this context. One response to the clear evidence that people's reasoning does not consistently conform to logical norms is the idea that there exist two separable inferential systems. One of these is meant to be a major source of variability in reasoning, while the other has at least the potential to reason more logically. Dual process theories (Sloman, 1996; Stanovich and West, 1998; Evans, 2007) postulate that people have two major inferential systems that interact with each other. Such theories have multiple forms and use different criteria to attempt to distinguish between these two systems. There is, unfortunately, no real consensus as to the characteristics or the definition of these two systems (see Evans and Stanovich, 2013, for a recent discussion). However, roughly speaking, these postulate a basically heuristic form of inference, which we is often referred to as System 1, which is presumed to be an evolutionarily primitive system that makes rapid inferences that are automated, contextual, use surface properties of problems, and rely extensively on stored knowledge. Such inferences are low-cost and do not involve working memory capacity. The second system, which postulates a more analytic form of inference, referred to as System 2, by contrast, is conscious, slow, and relatively costly in its use of working memory.

Although there are variable descriptions of the dual process framework, Evans and Stanovich (2013) suggest that one minimalist approach to this distinction is to suggest that heuristic inferences correspond to rapid, autonomous inferential processes, while analytic inferences are characterized by working memory-based processes that support hypothetical thinking. The latter are particularly characterized by cognitive decoupling, allowing inferences that are not necessarily tied to existing knowledge structures (Stanovich and Toplak, 2012). For our purposes, a key characteristic of System 1 inferences is they are intrinsically variable, since they necessarily reflect the idiosyncratic nature of people's internal representations. Given this intrinsic variability, there is no reason to think that a single normative model would ever be able to capture its properties, at least not within a relatively straightforward model. However, System 2 allows at least

the possibility of hypothetical thinking involving some degree of conscious processing of information. If this basic distinction is reasonably accurate, then the ability to even consider the possibility that normative models could exist must be the result of System 2 processing.

DEVELOPMENTAL EPISTEMOLOGY

With these distinctions in mind, there are two forms of argument for a normative model for System 2 reasoning. The first is an epistemological argument that partly owes its form to Piaget's work on cognitive development. Piaget in fact referred to his field of study as genetic epistemology. The basic argument, which was derived from Kant (see Henle, 1962) can be stated as follows. In order to adequately process the infinite variety of information that is potentially accessible, the human mind requires some basic categories. Kant assumed that the basic categories were a priori, that is, that they are a basic component of the human cognitive system and are essentially biological. Such an essentialist view of the basis of human cognition is actually quite common in many developmental theories. For example, studies examining people's understanding of categories and concepts (Gelman and Markman, 1986), and object permanence (Baillargeon et al., 1985) have indeed claimed that people have biological underpinnings that allow them to consistently extract specific categories or object qualities from complex forms of information. These are determined by the biological niche humans have constructed over evolutionary time. For example, understanding that a physical object retains its basic identity even when it changes shape or disappears would be a critical component of a cognitive system evolved to survive in a world in which there is are constantly moving objects. Similarly, and more in line with our current problem, understanding that objects that move by themselves can be considered to correspond to a category (which we refer to as implicitly living). Objects in such living categories are considered to have shared invisible attributes, which is a very useful way of conceptually dealing with the world in which separating out living from nonliving categories is a vital component, and understanding the specific properties of the latter can be particularly useful. Such a basically biologically based approach has in fact been proposed for human reasoning. As stated previously, Braine's natural logic approach (Braine, 1978) takes just such a stance. There is one problem with this, however. The distinction between living and nonliving categories, and the ability to understand primitive transformational consistencies, such as those required to understand object permanence are found in very young children, which at least suggests empirical support for a biological hypothesis. This is simply not the case for inferential reasoning, and the idea that there is an essential biological basis that corresponds to some form of internal rules of inference appears to be empirically untenable.

Piaget's approach to this problem was both biological and developmental (Piaget, 1971). In line with Kant, he assumed that the human mind did indeed require basic categories in order to adequately process information about an inordinately complex world. However, he did not assume that these categories were biological in origin. Instead, he postulated that biology provided the basic processes that allow systematic cognitive change. He also postulated that such changes were essentially systemic. In other words, he clearly made a distinction between the accumulation of knowledge, which could lead to a piecemeal and unconnected body of knowledge, and the development of the basic categories of mind that allowed people to process such knowledge. This latter can be considered the basis of the epistemology of the mind. In this perspective, the idea of a normative model of the mind can be seen as having the same basic function for cognition as the idea of a universal grammar has for language. More specifically, if human minds had essentially different epistemologies; that is, if they used different basic forms of categorization and reasoning, then the problem of just how people could communicate efficiently would arise.

Piaget added one component to this analysis. He started from empirical results that showed that children's understanding of the world appears to progress through different levels or stages. He early on remarked that young children appear to have a more primitive epistemology than adults that is the underlying basis of their thinking relied on basic categories that were less consistent than those that appear to underlie adult reasoning. For example, young children have variable notions of the basic concept of quantity, being unable to consider that quantity is an invariant property of mass (Piaget et al., 1997). The lack of such invariance makes their thinking inherently unstable. A parent who is faced with a child who does not want to eat a meal because there is too much food, and who mushes up the food into a smaller area, is using this instability to successfully manipulate his or her child. In contrast, adults have no problem understanding invariance of quantity, in fact most consider the questions used to examine this notion to be simply stupid, since the answers are self-evident.

It is this form of basic difference in epistemology that Piaget attempted to describe in his research program, something that has often been lost in the debates over details about the age at which specific abilities appear, etc. This approach thus assumed that there was change and development in such a basic epistemology. However, the course of this development is not really variable, but was meant to mirror both the physical and biological properties that are critical components of the basic cognitive system. Thus, Piaget postulated that there was an invariant developmental sequence by which the very primitive cognitive categories present at birth, combined with whatever innate tendencies might drive early information processing, would gradually transform into the adult version. Piaget also supposed that epistemological development would tend towards the same basic normative model. A critical point is understanding just why this would be true. One reason, which underpins Piaget's basic hypothesis is that epistemological structures are generated by interactions with the physical world. These start by interactions based on action and perception. Over time, and repeated interactions, children develop a logic of actions, which follows a coherent and nearly universal sequence (Piaget, 1965). In fact, this same sequence has been observed in primate species and some mammals (Scarr-Salapatek, 1976). Thus, it could already be argued that the end-point of sensori-motor development can be described by a normative model, one that reflects the deep structure of the way that humans structure physical action in the real world.

To make this distinction more explicit, when it is claimed that most 2 year-olds have the same basic epistemology, one that corresponds to a clear normative model, this does not mean that they have learned the same things. Specific learning clearly depends on the concrete environment in which children are raised. Thus a child might learn that cookies are good to eat in one context, and another might learn that candies are good to eat, depending on what is available. However, when trying to get cookies or candies or anything else that is desirable, that is hidden by an obstacle, all children will use the same action logic. Once again, the actual actions can vary (for example, one child might bat the obstacle away, while another might reach for it to move it away), but the logic of the action sequence is the same (action 1 is directed towards the obstacle with the aim of displacing it, action 2 is directed towards the goal). In fact, debates about sensori-motor cognition are mostly about whether sensori-motor logic is developed faster or slower, but there is little debate about the form of this logic. There are thus quite solid grounds for suggesting that development at this level can indeed be described by a normative model (Dasen, 1972).

Piaget's explanation of development considers that more abstract forms of conceptually based cognition are derived by processes of representation and symbolic manipulation of the logic of actions. Thus, for example, basic categorization is derived from the process of perceptually based similarity relations, causal categories are derived from direct causality, etc. More specifically, reasoning reflects the basic structure of conditional action schemas. For example, there are many ways to get a biscuit, thus understanding the uncertainty of such actions directly reflects what children have already learned about the physical world (Byrnes and Overton, 1986). Increasingly abstract forms of reasoning require a long and complex process of representation and restructuring of more abstract concepts, but these still reflect the underlying structure of the physical world. This point of view would thus suggest that the end-point of epistemological development would be essentially the same. Once again, if one examines studies looking at the ability of pre-adolescents to make inferences about the concrete world, empirical results strongly suggest that the same level is attained by most children, although there is a great variation in age. Thus, most children can understand that liquid quantities are conserved over transformations, allowing them to consistently infer that a simple change in container will not alter the quantity.

The same is found with transitive inferences, the logic of which most children understand by adolescence when the content is concrete and clear (Markovits et al., 1995). Once again, similarly to studies of sensori-motor development, debate about the form of such concrete logic is not about the form of such logic, but about whether the corresponding abilities are developed more or less rapidly. Thus, there is also very clear evidence that the logic of pre-adolescents, which allows understanding of many forms of conservation, and basic forms of transitivity, causality, categorization, etc. when these are applied to concrete, perceptible problems consistently described the abilities (and performance) of most pre-adolescents (although age of acquisition is highly variable). In other words, a normative model can be claimed to exist to describe the concrete reasoning of most children by adolescence. Since development on the more abstract, formal level is derived from reasoning structures developed previously, this suggests that formal reasoning should indeed correspond to a single form of normative model. This is indeed the underlying rationale for Piaget's claim that there is a single normative model for formal thinking.

COMMUNICATION AND NORMATIVE REASONING

Having a shared epistemology is certainly useful in order to allow different members of a species to process variable information in ways that are internally consistent. There is one further argument for the existence of a single normative model of human reasoning. If this normative model is a model of the underlying epistemology of the conscious component of the cognitive system, then the ability to communicate must require sharing the same basic epistemology. In fact, it could be argued that normalization of System 2 reasoning is particularly critical in this context. In order to understand this point, it is useful to consider communication with System 1 (intuitive) reasoning. Although much of just what is involved in such reasoning remains mysterious, we can fruitfully speculate about some aspects of this. Intuitive inferences can reflect (at least) two forms of information. The most critical of these from our point of view is the internal structure of experiences stored in memory. Theories such as probabilistic models of inference (Oaksford and Chater, 2003) assume that intuitive inferences reflect stored knowledge about the world. A useful example is the belief-bias effect, which is the tendency of people to accept conclusions that are judged as believable as being logically valid (Evans et al., 1983). Although experiments examining this effect use conclusions that are highly believable for a large number of people, believability is clearly personal. In addition, there is certainly an emotional component to the force with which believability acts on System 2 processing. One excellent example of this is given by a study by (Klaczynski, 2000). He found that interference with System 2 processing was much stronger when the beliefs that were being examined were related to a domain with very high emotional valence (religion) than when these were related to a domain with lower emotional valence (class). There are increasing numbers of studies that link emotional experiences to inference-making (Blanchette and Richards, 2010). Thus, it is reasonable to assume that Intuitive processing tends to be highly personal, in that it reflects idiosyncratic personal experience, which conditions not only understanding of the underlying structure of experience and events, but is complicated by emotional valences that clearly reflect individual experiences. Intuitive processing is by definition unconscious, and certainly experience shows us that idiosyncratic and emotionally driven inferences do not generate much in the way of metacognitive awareness.

In other words, intuitive processing can clearly serve individual purposes by allowing people to make rapid, low-cost inferences that reflect their past experience. Doing so increases the chances that future behavior will mirror past circumstances, which allows individuals to profit from experience in a very immediate way. If that were the end of the story, there would be no need for any other inferential system. However, there is another component to human behavior, and that is the fact that humans live in social

groups. This makes intercommunicability a critical component of any form of reasoning, since group behaviors require reconciling many divergent individual agendas in a way that allows group cohesion. Further complicating this dynamic is that fact that group complexity increases exponentially with numbers of group members. This explosion of social information has been hypothesized to be a major evolutionary driver for human cognition (Dunbar, 1993). Recently, Mercier (Mercier, 2011; Mercier and Sperber, 2011) has proposed a general evolutionary theory for the development of System 2 reasoning abilities that suggest that these abilities have evolved in order to regulate communication in complex social groups.

This perspective views reasoning as a form of argumentation which allows people both to present overt reasons for actions in an attempt to convince others and to allow others to evaluate arguments. While reasoning might have other functions, it is a useful hypothesis to see it as a means for exchanging explicit reasons for action, since this would have the potential to allow groups to make decisions that were more efficient than a simple majority rule would provide. Seeing reasoning as a form of communication, or even as a useful underpinning for communication, makes an even stronger case for the existence of a common epistemological core, since the essence of argumentation requires a sufficiently strong common basis.

Now, there are (at least) two ways that effective communication can be insured. The first involves the use of common biologically based intuitive schemas. In the absence of explicit language or symbolic thought, such schemas characterize the social cohesion of many social animals. There is reasonable evidence that humans also have such implicit social schemas. For example, we have recently shown that humans share an intuition about coalition formation as a function of individual power that is similar to what has been found behaviorally in chimpanzees (Benenson et al., 2009). Another form of implicit inference is that underlying the Gricean view of linguistic communication, in which pragmatic interpretations of language acts are underpinned by common assumptions that are derived from shared experience (Grice, 1981). However, such intuitive schemas can only work well when social behavior is relatively constrained. Human social behavior, while certainly sharing many aspects with more biologically constrained social species, is, however, very flexible. Flexibility, while allowing greater ability to adapt to changing circumstances, has a clear effect on the possibility of ensuring effective communication solely on the basis of shared intuitive schemas. This puts a greater functional burden on overt, language based communication to ensure that social interactions do not degenerate into conflicts based on differing individual intuitions and perceptions. But, in order for such communication to serve this function, it is imperative that there exist a shared epistemology, i.e., that the ideas shared overtly are underpinned by some common basic principles of reasoning. This again provides theoretical weight for the idea that explicit human reasoning should have a normative core.

WHAT WOULD A NORMATIVE MODEL OF REASONING LOOK LIKE?

There are thus reasons related to the basic epistemology of human cognition and to the importance of reasoning to communication that suggest the necessity of a single normative model for explicit human reasoning. There are some basic considerations that can give clues to just what this model entails. The first is an argument from conceptual power. The last couple of hundred years have led to a proliferation of models of logic that have radically different underpinnings. Each of these models is a product of the human mind, individually or in concert with others. One important constraint for a normative model that does indeed correspond to the workings of the explicit, analytic mind is that it should allow the construction of multiple forms of models of logic. In addition, such a model should be readily understandable by most people, exactly because of the premise of communicability that we have claimed previously. Both of these constraints, along with historical and developmental considerations suggest that the normative model of the mind should involve some basic principles that underpin the notion of validity.

Secondly, one of the key components of cognitive development is that change goes towards increased complexity. A good example of this is well-documented in language learning, the phenomenon of over-generalization. Young children are faced with a complex variety of linguistic forms, with many idiosyncratic formulations which have historical roots, but often violate what are more frequent forms. Children's strategies for learning language is to identify (by whatever process this is done) the most frequent pattern, and generalize this as a rule that is used in all occurrences of a given class, even when this involves generating words that have never actually been encountered (Onnis et al., 2002). Young children do the same with cognitive categories, picking out simple rules and extending these to concepts that are not actually instances of these categories. In other words, the developing human mind has a clear strategy, which requires generating simple rules and extending these to a wide variety of instances. It is only through continued interactions and reflection that these initial simple rules are extended to more complex concepts. What this in turn suggests is that if basic analytic reasoning relies on core normative principles, it will take a form that reduces the cognitive load required to reason. Thus, while there are a multitude of logical systems that take into account the true complexity of human experience, this argument suggests that normative principles will be less complex than any of these logics.

In this perspective, developmental studies can provide some very useful information. As we have seen, variability is a key characteristic of the reasoning of adults. However, while such variability is suggestive of a system of thought that has no common epistemology, the addition of an intuitive component of the functioning of the human mind makes variability more readily explicable in terms of the combination of a personal form of intuition which reflects individual experience and explicit, analytic thought that despite some variability has similar underpinnings.

The problem here is to determine exactly what are the critical components of normative thinking. The key to this is to distinguish between a normative model that is determined by its outcome, and one that is described by the nature of the underlying processes. Most studies that have looked at whether people reason normatively have examined outcomes, specifically whether the responses given to inferential problems are the same as whatever norm is being compared to. However, unless it is

believed that inferential rules are constructed directly in the mind (as some theories do indeed suggest, e.g., Braine, 1978; Rips, 1983), we can rephrase the question of normality by trying to specify what kinds of underlying analyses must be implied by any system of thought that can in principle produce "correct" answers.

CHILDREN'S UNDERSTANDING OF VALIDITY

In this perspective, we can distinguish at least two major components of basic logic that are necessary to produce a form of reasoning that can in principle become powerful enough to generate complex normative models. One important use of such a logic is the ability to explicitly examine the consequences of intuitive forms of reasoning in a way that allows people who do not share the same intuition to communicate. One key component of this ability is understanding the distinction between belief and some form of validity. In other words, before people can explicitly examine and compare the consequences of divergent personal experience, they must be able to distinguish, at least in principle, inferences that are derived directly from experience and those derived by some process of "logical reasoning".

If this corresponds to a basic component of human reasoning, then it should be evident, in some form, in children. In fact, there is clear evidence that this distinction is fairly primitive. For example, Moshman and Franks (1986) found a clear developmental trend so that by early adolescence, most children can spontaneously understand the distinction between belief and validity, well before the level of schooling in which this distinction is taught. More strikingly, Morris (2000) found that children as young as 5-years of age, when given an appropriate content can generate this distinction. In other words, understanding the distinction between belief and validity is an early developmental acquisition, a critical one if explicit reasoning is indeed a counterpoint to intuitive reasoning. Of course, this is not really news to parents of young children, who despite the real difficulties of doing so, are nonetheless able to "reason" with children in a way that confronts the child's intuitions with some form of logic.

POSSIBILITY AND NECESSITY: COUNTEREXAMPLES IN REASONING

If children can understand the distinction between validity and belief, the next question is just how validity is determined. This of course goes to the question of just what kind of "logic" is available to children, and how this is related to the logic of adults. I have claimed (Markovits, 1993) that the key component of understanding this question derives from one of Piaget's later works, on the relation between possibility and necessity (Piaget, 1987).

Before examining this, it is useful to make an important distinction underlying Piaget's approach. Piaget proposed standard propositional logic as a competence model for advanced adult reasoning. This has often been interpreted as implying a rule-based form of reasoning, which would invariably lead to standard logical responses. However, this is a mischaracterization. What was specifically proposed was that the underlying epistemology that characterized advanced adult reasoning would allow the ability to generate such responses. The work on possibility and necessity was an attempt to specify the nature of this epistemology. The basic question that was raised concerns the kinds of factors that can explain how children and adults can conclude that a potential conclusion is necessary. The analysis makes no mention of rules, instead it places such logical conclusions within the more general context of the range of information that can be generated by the reasoner. In this, one critical component of reasoning is the range of possibilities that are processed by children in the context of a given problem. A given inference is necessary if it excludes whatever possibilities are generated by the child at the moment of reasoning. This corresponds to what one could refer to as local necessity, since the actual generality of a conclusion depends critically on the range of possibilities that are generated. Critically, if a child or adult is aware of a possibility that is not excluded, they will reject a given conclusion. This interaction between possibilities and necessity is modulated by the degree of abstraction of the processes used to analyze a given problem.

A similar idea underlies the mental model analysis of reasoning (Johnson-Laird, 2001; Johnson-Laird and Byrne, 2002). Mental model theory considers that people construct internal representations of possible states of the world (models) that characterize the major premise of a given inference. Possible states must be generated by a reasoner, based on combinations of semantic and pragmatic factors (Byrne, 2005). Critically, an inference is considered to be valid if there are no explicit counterexamples in the reasoner's representation. The presence of a counterexample is sufficient to render a putative conclusion invalid. An inferential judgment is thus an on-line consequence of a reasoner's ability to (1) generate a more or less full range of possibilities consistent with premises and (2) determine whether these possibilities contain a counterexample.

One important distinction between the Piagetian analysis and mental model theory is that the latter postulates that internal representations are derived from a semantic analysis of logical connectors, albeit one that is modulated by pragmatics. Since the semantics of logical connectors are assumed to be generally invariant, variation in reasoning performance is accounted for by such individual difference factors as memory capacity. Mental model theory does not have a very clear developmental component, although Barrouillet and colleagues (Barrouillet and Lecas, 1999; Barrouillet et al., 2008) have suggested that working memory limitations can affect children's ability to actually represent the full semantics of logical connectors. Development will necessarily tend towards the same forms of logical reasoning that are determined by the shared semantics of logical connectors. Thus, although mental model theory presents an analysis of reasoning that is consistent with the interaction between possibility and necessity, the developmental component has a very different focus (see Markovits and Barrouillet, 2002 for a version of mental model theory that has a developmental focus that is more consistent with the Piagetian model).

Piaget's work indicated that one of the key factors in development is the ability to generate increasingly abstract forms of possibilities (Gauffroy and Barrouillet, 2011). Young children start by considering possibilities that are more concrete and related to the situational factors for a given problem context. With development, these possibilities become more extended and abstract, and less tied to situational constraints. However, by 6- or 7-years

of age (Markovits, 2000), children can understand that a conclusion that eliminates all other possibilities is necessary. Of course, since this judgment of necessity depends on the range of possibilities that are eliminated when this conclusion is made, it is subject to revision even if exactly the same form of reasoning is applied, since the generation of possibilities can vary from one moment to the next. In other words, reasoning at this level is sequentially defeasible (Pollock, 1987), that is the same person can arrive at a different conclusion for the same inference, simply because the domain of possibilities accessed during reasoning might change, see Markovits (1985) and Byrne (1989) for examples of defeasible reasoning in adults. However, the *processes* by which judgments of validity are made are in principle general. More importantly, it is possible to overtly challenge any such inference by comparing possibilities. This in turn is realistic only because judgments of validity depend, not on an accumulation of data, but on the presence or absence of a counter-example to a given conclusion.

There is one further point that can be made in this context. As noted previously, it has been argued that most relations are, in real life probabilistic (Oaksford and Chater, 2007). If the point of reasoning was to faithfully reflect the real characteristics of the real world, then one would expect reasoning to be essentially probabilistic. However, the analysis of "logical" reasoning that I use here does not generate conclusions such as "it is probably true that. . .." It simply allows judgments of validity or not, in other words, it allows concluding that something is certain or that it is not. Why should this form of judgment be useful when trying to reasoning about phenomena that are inherently probabilistic? There are some good reasons for this, but the chief one is that it is cognitively much simpler. While a probabilistic judgment or any other kind of intuitive judgment that relies on stored knowledge about the world can conceivably be made very rapidly by associative processes, explicitly communicating the basis for such judgments would in theory require explicitly processing a large quantity of information. In fact, in many cases most people are unable to do such explicit processing. In this, case comparing conclusions would simply result in people fighting over personal beliefs. Even if we assume that a reasoner does (remarkably) have conscious access to the relevant information, and is consciously aware that conclusion X is probable because of data set Y, the problem of how someone else, whose personal data base does not contain the same information, can process this conclusion.

A useful, because somewhat more real example, can be taken from research on aggressive behavior. One of the underlying mechanisms of such behavior concerns the expectation (mental model) that a given social interaction will have an aggressive outcome (Dodge et al., 1990). Children who develop a strong expectation that interactions will be aggressive, tend to infer that most actions will have an aggressive outcome. Now, imagine that two children with differing expectations are interacting, and attempting to determine the outcome of a given course of action. In order to make a reasonable comparison, both children would have to recount the many kinds of interactions that form the basis of their expectations. Then each would have to attempt to integrate the other's information into their own internal data base, a daunting task at best, and one that would require not only a great deal more cognitive resources that most children possess, but a great deal of time. The fact that such discussions do not really take place might well account for the difficulty in lowering aggressive behavior (Dishion et al., 1999). In contrast, since many arguments consist in deciding on taking a given course of action or not, a simple counterexample strategy would be quick and useful, precisely because of its inherent limitations. If this is indeed characteristic of basic reasoning, then once again, developmental studies should provide evidence.

THE DEVELOPMENT OF ELEMENTARY CONDITIONAL REASONING

We can illustrate this evidence by examining simple conditional (if–then) reasoning. Conditional reasoning is one of the key components of propositional logic, and is certainly one, if not the most, frequently studies forms of reasoning. Piaget considered that conditional reasoning was one of the key competencies of formal operational thinking. The basic approach that we are taking here suggests that understanding the development of conditional reasoning will rest upon the basic understanding of the interplay between possibility and necessity and the way that this can be used increasingly abstract context. The key component of this suggestion is identification of just what the range of possibilities are required in order to make "logical" conditional inferences, and whether children are able to process these in a way that allows them to make simple judgments of validity.

To simplify this discussion, we focus on two conditional inferences. The Modus ponens (MP) inference involves reasoning that "If P then Q, P is true." The affirmation of the consequent (AC) inference is "If P then Q. Q is true." The logical conclusion to a MP inference is that "Q is true". Now studies with adults have shown that the ability to conclude that the MP inference is valid depends on the extent to which disabling conditions (Cummins et al., 1991; Cummins, 1995) are incorporated into people's representations of premises. A disabling condition (disabler) is a condition that is associated with a given P then Q premise, one that could invalid the link between antecedent and consequent terms. The strongest form of disabler is given when reasoning with empirically false premises, for which the true relation is a disabler. Very young children have little problem in considering MP inferences to be valid for a variety of contents. The one major exception is reasoning with empirically false premises, for which young children have great difficulties in accepting the MP inference. However, interventions that allow them to inhibit retrieval of the true disabler, such as embedding premises into a fantasy context (Dias and Harris, 1988, 1990) dramatically improve their ability to accept this inference. With increasing age, children are more able to spontaneously inhibit potential disablers when given standard logical instructions (Markovits and Vachon, 1989).

In contrast, there is no logical conclusion to an AC inference, although typically, many people respond to AC inferences by concluding that "P is true" (Cummins et al., 1991; Thompson, 1994). The reason for the theoretical lack of any logical conclusion is that a true conditional allows for the possibility of what have been referred to as alternative antecedents, that is cases where alternate conditionals "If A then Q" are, or might be true. Considering such alternatives (which comprise the domain of possibilities in

this case), allows producing a counterexample to what is the usual AC conclusion. In fact, there is strong empirical evidence that the availability of such alternative antecedents in memory is a strong determinant of the conclusions given to AC inferences (Markovits and Vachon, 1990; Cummins et al., 1991; Thompson, 1994; Quinn and Markovits, 1998). In other words, when people can access such alternatives when reasoning, they tend to reject the certain conclusion for AC inferences.

Analysis of both the MP and the AC inference shows that, in both cases, children's judgments of validity are determined by the kind of information that is incorporated into the representation of premises, during the online process of reasoning. Although the specific information varies (disablers and/or alternative antecedents), the basic dynamic is the same. Even young children will accept a conclusion as valid if their representation of the premises does not include a potential counterexample, otherwise the conclusion will be considered to be invalid. Finally, although considering these two inferences is particularly useful, it is worth briefly examining the two other inferences that define conditional logic. Studies have generally shown that responses to the denial of the antecedent (DA) inference responds to the presence or the absence of alternative antecedents in the same way as AC inferences. The developmental pattern is also very similar. Similarly, the modus tollens (MT) inference corresponds somewhat similarly to the MP inference to the presence of disablers. However, the use of negation in both these inferences somewhat complicates analysis. Understanding the basic notions underlying the understanding of logical validity is more easily presented by concentrating on the MP and the AC inferences.

I have in fact argued that the best elementary definition of "logical" reasoning is the ability to understand the certainty of the MP inference and to simultaneously understand the uncertainty of the AC inference (Markovits and Lortie Forgues, 2011). As I have stated, the key factor in this form of reasoning is the kind of information that is incorporated into children's representations of premises. Constructing a representation of premises that does not include potential disablers, but that does include potential alternative antecedents requires inhibiting the former while retrieving the latter. The tension between these two contradictory cognitive processes can explain why even educated adults find it difficult to reason logically in the limited sense that we use (see Markovits, 2014). However, there is strong empirical evidence that children as young as 6–7-years of age are indeed able to understand both the certainty of the MP inference and the uncertainty of the AC inference, when the content of conditional premises allow for very ready access to alternatives. For example, when reasoning with propositions such as "If an animal is a dog, then it has four legs. An animal has four legs. Is it a dog?," young children will readily reject this inference (Markovits, 2000). More tellingly, they will do so by explicitly citing the existence of a counterexample to the implied conclusion ("cats have four legs"). When given problems where the alternatives are explicitly presented, even younger children will reject the AC inference while accepting the MP inference (Markovits and Thompson, 2008).

In other words, very young children are able to reason "logically" with simple, direct forms of propositional logic when the content allows appropriate inhibition and retrieval of relevant information. Given the strong tendency of children to accept inferences, the most striking part of these results is that they are capable of processing the fact that there are alternatives to a putative conclusion, and using this as a basis for rejecting a putative conclusion. I would then argue that this is exactly the nature of the kind of logical reasoning that is the "norm" for most people, which involves constructing a simple representation of an inference and deciding on the validity of a potential conclusion based on the existence of a possible counterexample.

Does this model imply that children or even educated adults will consistently give the standard propositional logic answer to (for example) all AC inferences. Not at all. Since the exact response to any inference depends on the range of counterexamples that are generated while reasoning, which is in turn related to retrieval and inhibitory processes, variability is expected. This is an important point to make. The increasing ability to generate potential counterexamples can in fact produce variable responses. For example, children who are more efficient in simply producing potential causal alternatives tend to reject the MP inference more frequently (Janveau-Brennan and Markovits, 1999). This corresponds to the notion of local necessity, in which validity is the result of application of the basic principle related to counterexample detection to an online process of possibility generation. Thus, this model makes no presuppositions about the nature of the underlying rules or algorithms. Such a basic epistemology is thus potentially consistent with a wide variety of reasoning systems, including for example the incomplete logic proposed by Gauffroy and Barrouillet (2009).

The important part of this conception of reasoning is the idea that people can recognize the presence or the absence of a counterexample and use this to make a judgment of validity that is internally consistent. Critically, especially for its usefulness in communication, people can adjust their inferences based on externally communicated counterexamples. Evidence shows that both children (Rumain et al., 1983) and adults (Markovits, 1985) will revise conclusions to AC inferences when given additional information about potential alternative antecedents. Similarly, explicit information suggesting disablers results in adults rejecting the MP inference (Byrne, 1989). The effect of providing explicit information about the existence of potential counterexamples is to reliably increase rates of rejection of conclusions that were previously accepted. In other words, both children and adults are able to revise conclusions that are generated by their internal inferential processes simply by considering alternatives furnished by other people. In a similar vein, Klaczynski (2001) has shown that children are able to recognize logical arguments that rely on the presence of counterexamples as being superior to arguments that rely on other kinds of processes.

In this context, basic logical reasoning can be seen as the ability to use a counterexample to invalidate an otherwise acceptable conclusion. Empirical results show that this ability is present at a very early age. Critically, there is also evidence that both children and adults can use externally presented counterexamples to modify their own inferences. Thus logical reasoning has both an internal function that allows for judgments of validity and an external function that provides a basis for modifying personal judgments by considering specific arguments that present potential counterexamples.

If such an ability does indeed represent the kernel of logical reasoning, then development should reflect not any change in this basic form of logic, but should be tied to the increasing ability to generate and inhibit potential counterexamples in an increasingly abstract way. I have indeed argued that this is a good description of the development of simple conditional reasoning between early childhood and later adolescence (Markovits, 2013). To summarize this pattern, young children can consistently reject the AC premise and accept the MP premise for premises that use if–then relations that link classes and properties (Markovits, 2000). The ability to do the same with causal conditionals ("If cause P then effect Q") is found only in pre-adolescents (Janveau-Brennan and Markovits, 1999). Reasoning this way with contrary-to-fact premises is a later development (Markovits and Vachon, 1989). Finally, understanding the certainty of the MP inference and the uncertainty of the AC inference with completely abstract premises is a much later development, and is only found with educated adults (Venet and Markovits, 2001; Markovits and Lortie Forgues, 2011). In other words, the developmental pattern is completely consistent with the idea that the basic mechanism underlying simple logical reasoning is available to quite young children, and that subsequent development extends this same process to increasingly complex and abstract forms of content.

Importantly, extending basic reasoning to abstract content (Markovits and Lortie Forgues, 2011) gives the ability to reason logically even with content that has no concrete referents or for which the referents are unintuitive. Such a form of reasoning is the historical basis for the various models of logic that have been constructed, and which correspond to a wide variety of different forms of reasoning. In other words, the kind of abstract reasoning that develops from the primitive base found in young children can become powerful enough to allow people to construct normative models of many types.

REFERENCES

Baillargeon, R., Spelke, E. S., and Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition* 20, 191–208. doi: 10.1016/0010-0277(85)90008-3

Barrouillet, P., Gauffroy, C., and Lecas, J.-F. O. (2008). Mental models and the suppositional account of conditionals. *Psychol. Rev.* 115, 760–771. doi: 10.1037/0033-295X.115.3.760

Barrouillet, P., and Lecas, J.-F. (1999). Mental models in conditional reasoning and working memory. *Think. Reason.* 5, 289–302. doi: 10.1080/135467899393940

Benenson, J., Markovits, H., Thompson, M. E., and Wrangham, R. W. (2009). Strength determines coalitional strategies in humans. *Proc. Roy. Soc. B Biol. Sci.* 276, 2589–2595. doi: 10.1098/rspb.2009.0314

Blanchette, I., and Richards, A. (2010). The influence of affect on higher level cognition: a review of research on interpretation, judgement, decision making and reasoning. *Cogn. Emot.* 24, 561–595. doi: 10.1080/02699930903132496

Braine, M. D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychol. Rev.* 85, 1–21. doi: 10.1037/0033-295X.85.1.1

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition* 31, 61–83. doi: 10.1016/0010-0277(89)90018-8

Byrne, R. M. J. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

Byrnes, J. P., and Overton, W. F. (1986). Reasoning about certainty and uncertainty in concrete, causal, and propositional contexts. *Dev. Psychol.* 22, 793. doi: 10.1037/0012-1649.22.6.793

Cummins, D. D. (1995). Naive theories and causal deduction. *Mem. Cogn.* 23, 646–658. doi: 10.3758/BF03197265

Cummins, D. D., Lubart, T., Alksnis, O., and Rist, R. (1991). Conditional reasoning and causation. *Mem. Cogn.* 19, 274–282. doi: 10.3758/BF03211151

Dasen, P. R. (1972). Cross-cultural Piagetian research: a summary. *J. Cross Cult. Psychol.* 3, 23–40. doi: 10.1177/002202217200300102

Dias, M. G., and Harris, P. L. (1988). The effect of make-believe play on deductive reasoning. *Brit. J. Dev. Psychol.* 6, 207–221. doi: 10.1111/j.2044-835X.1988.tb01095.x

Dias, M. G., and Harris, P. L. (1990). The influence of the imagination on reasoning by young children. *Brit. J. Dev. Psychol.* 8, 305–318. doi: 10.1111/j.2044-835X.1990.tb00847.x

Dishion, T. J., McCord, J., and Poulin, F. (1999). When interventions harm: peer groups and problem behavior. *Am. Psychol.* 54, 755–764. doi: 10.1037/0003-066X.54.9.755

Dodge, K. A., Bates, J. E., and Pettit, G. S. (1990). Mechanisms in the cycle of violence. *Science* 250, 1678–1683. doi: 10.1126/science.2270481

Dunbar, R. I. (1993). Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci.* 16, 681–693. doi: 10.1017/S0140525X00032325

Elqayam, S., and Evans, J. S. B. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Elqayam, S., and Evans, J. S. B. (2013). Rationality in the new paradigm: strict versus soft Bayesian approaches. *Think. Reason.* 19, 453–470. doi: 10.1080/13546783.2013.834268

Evans, J. S. B. T. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. New York, NY: Psychology Press.

Evans, J. S. B. T., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Mem. Cogn.* 11, 295–306. doi: 10.3758/BF03196976

Evans, J. S. B. T., Over, D. E., and Handley, S. J. (2007). "Rethinking the model theory of conditionals," in *The Mental Models Theory of Reasoning: Refinements and Extensions*, eds W. Schaeken, A. Schroyens, and G. d'Ydewalle (Mahwah, NJ: Lawrence Erlbaum Associates), 63–83.

Evans, J. S. B., and Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Gauffroy, C., and Barrouillet, P. (2009). Heuristic and analytic processes in mental models for conditionals: an integrative developmental theory. *Dev. Rev.* 29, 249–282. doi: 10.1016/j.dr.2009.09.002

Gauffroy, C., and Barrouillet, P. (2011). The primacy of thinking about possibilities in the development of reasoning. *Dev. Psychol.* 47, 1000–1011. doi: 10.1037/a0023269

Gelman, S. A., and Markman, E. M. (1986). Categories and induction in young children. *Cognition* 23, 183–209. doi: 10.1016/0010-0277(86)90034-X

Gigerenzer, G., and Selten, R. (2001). "The adaptive toolbox," in *Bounded Rationality: The Adaptive Toolbox*, eds G. Gigerenzer and R. Selten (Cambridge, MA: MIT Press), 37–50.

Grice, H. P. (1981). "Presupposition and conversational implicature," in *Syntax and Semantics*, Vol. 9, *Pragmatics*, ed. P. Cole (New York, NY: Academic Press), 183–198.

Henle, M. (1962). On the relation between logic, and thinking. *Psychol. Rev.* 69, 366–378. doi: 10.1037/h0042043

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books. doi: 10.1037/10034-000

Janveau-Brennan, G., and Markovits, H. (1999). The development of reasoning with causal conditionals. *Dev. Psychol.* 35, 904–911. doi: 10.1037/0012-1649.35.4.904

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends Cogn. Sci.* 5, 434–442. doi: 10.1016/S1364-6613(00)01751-4

Johnson-Laird, P. N., and Byrne, R. M. J. (2002). Conditionals: a theory of meaning, pragmatics and inference. *Psychol. Rev.* 109, 646–678. doi: 10.1037/0033-295X.109.4.646

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: a two-process approach to adolescent cognition. *Child Dev.* 71, 1347–1366. doi: 10.1111/1467-8624.00232

Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision-making. *Child Dev.* 72, 844–861. doi: 10.1111/1467-8624.00319

Markovits, H. (1985). Incorrect conditional reasoning among adults: competence or performance? *Br. J. Psychol.* 76, 241–247. doi: 10.1111/j.2044-8295.1985.tb01948.x

Markovits, H. (1993). The development of conditional reasoning: a Piagetian reformulation of mental models theory. *Merrill Palmer Q.* 39, 131–158.

Markovits, H. (2000). A mental model analysis of young children's conditional reasoning with meaningful premises. *Think. Reason.* 6, 335–347. doi: 10.1080/135467800750038166

Markovits, H. (2013). *The Developmental Psychology of Reasoning and Decision-Making*. New York, NY: Psychology Press.

Markovits, H. (2014). "Conditional reasoning and semantic memory retrieval," in *Reasoning and memory*, eds A. Feeney and V. Thompson (Hove UK: Psychology Press).

Markovits, H., Dumas, C., and Malfait, N. (1995). Understanding transitivity of a spatial relationship: a developmental analysis. *J. Exp. Child Psychol.* 59, 124–141. doi: 10.1006/jecp.1995.1005

Markovits, H. and Barrouillet, P. (2002). The development of conditional reasoning: a mental model account. *Dev. Rev.* 22, 5–36. doi: 10.1006/drev.2000.0533

Markovits, H., and Lortie Forgues, H. (2011). Conditional reasoning with false premises facilitates the transition between familiar and abstract reasoning. *Child Dev.* 82, 646–660. doi: 10.1111/j.1467-8624.2010.01526.x

Markovits, H., and Thompson, V. (2008). Different developmental patterns of simple deductive and probabilistic inferential reasoning. *Mem. Cogn.* 36, 1066–1078. doi: 10.3758/MC.36.6.1066

Markovits, H., and Vachon, R. (1989). Reasoning with contrary-to-fact propositions. *J. Exp. Child Psychol.* 47, 398–412. doi: 10.1016/0022-0965(89)90021-0

Markovits, H., and Vachon, R. (1990). Conditional reasoning, representation, and level of abstraction. *Dev. Psychol.* 26, 942–951. doi: 10.1037/0012-1649.26.6.942

Mercier, H. (2011). On the universality of argumentative reasoning. *J. Cogn. Cult.* 11, 1–2. doi: 10.1163/156853711X568707

Mercier, H., and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–74. doi: 10.1017/S0140525X10000968

Morris, A. K. (2000). Development of logical reasoning: children's ability to verbally explain the nature of the distinction between logical and nonlogical forms of argument. *Dev. Psychol.* 36, 741–758. doi: 10.1037/0012-1649.36.6.741

Moshman, D., and Franks, B. A. (1986). Development of the concept of inferential validity. *Child Dev.* 57, 153–165. doi: 10.2307/1130647

Oaksford, M., and Chater, N. (2003). Conditional probability and the cognitive science of conditional reasoning. *Mind Lang.* 18, 359–379. doi: 10.1111/1468-0017.00232

Oaksford, M., and Chater, N. (2007). *Baysian Rationality*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198524496.001.0001

Oaksford, M., and Chater, N. (2012). Dual processes, probabilities, and cognitive architecture. *Mind Soc.* 11, 15–26. doi: 10.1007/s11299-011-0096-3

Oaksford, M., and Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Think. Reason.* 19, 346–379. doi: 10.1080/13546783.2013.808163

Onnis, L., Roberts, M., and Chater, N. (2002). Simplicity: a cure for overgeneralizations in language acquisition? Contexts, 1, C2.

Overton, W. F., Ward, S. L., Black, J., Noveck, I. A., and O'Brien, D. P. (1987). Form and content in the development of deductive reasoning. *Dev. Psychol.* 23, 22–30. doi: 10.1037/0012-1649.23.1.22

Piaget, J. (1965). "The stages of the intellectual development of the child" in *Educational Psychology in Context*, eds B. A. Marlowe and A. S. Canestrari (Thousand Oaks, CA: Sage Publications), 98–106.

Piaget, J. (1971). *Biology and Knowledge: An Essay on the Relations between Organic Regulations and Cognitive Processes*. Oxford: University of Chicago Press.

Piaget, J. (1987). *Possibility and Necessity*. Vol. 1*: The Role of Possibility in Cognitive Development*, trans. H. Feider (Minneapolis, MN: University of Minnesota Press).

Piaget, J., Inhelder, B., and Pomerans, A. (1997). *The Child's Construction of Quantities: Conservation and Atomism*. London: Routledge.

Pollock, J. L. (1987). Defeasible reasoning. *Cogn. Sci.* 11, 481–518. doi: 10.1207/s15516709cog1104_4

Quinn, S., and Markovits, H. (1998). Conditional reasoning, causality, and the structure of semantic memory: strength of association as a predictive factor for content effects. *Cognition* 68, B93–B101. doi: 10.1016/S0010-0277(98)00053-5

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychol. Rev.* 90, 38–71. doi: 10.1037/0033-295X.90.1.38

Rumain, B., Connell, J., and Braine, M. D. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: if is not the biconditional. *Dev. Psychol.* 19, 471–481. doi: 10.1037/0012-1649.19.4.471

Scarr-Salapatek, S. (1976). "An evolutionary perspective on infant intelligence: species patterns and individual variations" in *Origins of Intelligence*, ed. M. Lewis (New York, NY: Plenum Press), 165–197. doi: 10.1007/978-1-4684-6961-5_6

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 119, 3–22. doi: 10.1037/0033-2909.119.1.3

Stanovich, K. E., and Toplak, M. E. (2012). Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind Soc.* 11, 3–13. doi: 10.1007/s11299-011-0093-6

Stanovich, K. E., and West, R. F. (1998). Individual differences in rational thought. *J. Exp. Psychol. Gen.* 127, 161–188. doi: 10.1037/0096-3445.127.2.161

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn. Sci.* 10, 309–318. doi: 10.1016/j.tics.2006.05.009

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Mem. Cogn.* 22, 742–758. doi: 10.3758/BF03209259

Tversky, A., and Kahneman, D. (2004). *Judgment Under Uncertainty: Heuristics and Biases*. New York, NY: Psychology Press.

Venet, M., and Markovits, H. (2001). Understanding uncertainty with abstract conditional premises. *Merrill Palmer Q.* 47, 74–99. doi: 10.1353/mpq.2001.0006

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 February 2014; accepted: 05 May 2014; published online: 23 May 2014.
*Citation: Markovits H (2014) Development and necessary norms of reasoning. Front. Psychol. **5**:488. doi: 10.3389/fpsyg.2014.00488*
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

OPINION ARTICLE
published: 30 January 2014
doi: 10.3389/fpsyg.2014.00049

In search for a standard of rationality

Emmanuel M. Pothos[1]* and Jerome R. Busemeyer[2]

[1] Department of Psychology, City University London, London, UK
[2] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA
*Correspondence: e.m.pothos@gmail.com

Edited by:
Shira Elqayam, De Montfort University, UK

Keywords: decision making, classical probability theory, quantum probability theory, rationality, conjunction fallacy

The debate on human rationality goes to the heart of fundamental questions about human existence. When can we say that a decision is correct? What is the basis for the achievements of the human intellect? What is the most important cognitive distinction between humans and non-human organisms? Proposals of rationality have an interesting status as psychological theories. They are not quite theories of decision making in practice—such theories are referred to as descriptive theories, to imply that they describe what goes on. Rather, proposals of rationality are normative theories, to imply theories of how people ought to reason, if they seek decision outcomes, which are deemed to be correct, on the basis of some absolute standard (here, we are simplifying a complex debate; arguments have been expressed against a distinction between normative and descriptive rationality as we make above, e.g., Elqayam and Evans, 2011, 2013). Of course, a normative theory must be partly a descriptive theory as well, since it is assumed that humans can, in principle, sometimes, reason on the basis of the normative prescription (they may just not do so, in typical situations, perhaps due to process demands or time or other constraints).

Currently, the dominant approach to normative rationality is based on classical probability (CP) theory. This approach was established after a major shift in conceptual thinking about rationality. Before (effectively since antiquity), it was believed that the standard for correctness in decision making was classical logic. But this position came under intense scrutiny, with experimental results showing that naïve participants, even in simple tasks, would not reason in a way consistent with classical logic (Wason, 1960). One reaction to such results was to develop dual theories of reasoning, which would, broadly speaking, involve a rational component and a heuristics one (cf. Sloman, 1996). However, a priori arguments emerged against any kind of role of logic in human practical decision making (i.e., decision making exempting mathematical/ scientific etc.; Chater and Oaksford, 1993). By contrast, a theory of rationality (and decision making) based on probability theory seemed (and seems) to align itself closely with the intuition we have about what it means for decision making to be successful. Such a theory is about the use of available information from the environment, so as to optimally predict the probability of future events (Anderson, 1991; Oaksford and Chater, 2007; Tenenbaum et al., 2011). The fact that the classical prescription appears to be consistent with human cognition in many cases (e.g., Griffiths et al., 2010, and the above references) also corroborates the psychological relevance of CP theory.

There are formal arguments to support the notion that CP theory provides a *correct* association of probabilities to uncertain events. The Dutch Book Theorem (DBT; e.g., Howson and Urbach, 1993) shows that if one assigns probabilities to events in a way inconsistent with the axioms of CP theory, then it is possible to identify a combination of stakes (money to be won or lost, depending on whether the events occur or not), which guarantees a loss (or gain, depending on the sign of the stakes). That is, according to the DBT, when failing to follow the rules of CP theory, you may be vulnerable to a sure loss (extensions to the DBT, such as the Converse DBT, have been presented too; Vineberg, 2011). Note that the DBT is based on value maximization, but it is well established that reasoners are typically e.g., risk averse (Kahneman and Tversky, 1979). Wakker (2010) showed that risk averse decision makers are subject to a Dutch Book, which provides an interesting conundrum, since expected utility theory, which allows for a risk averse utility function, is considered the rational theory of risky decision making. Nevertheless, the utility of the DBT, in relation to a theory of rationality based on CP theory, is that it provides a formal justification for why CP theory provides the normative prescription for decision making. In other words, currently, if one is interested in whether a probabilistic decision is correct or not, then one needs to explore its consistency with the prescription from CP theory.

The above is an extremely powerful and useful conclusion. Unfortunately, we believe it is vulnerable to criticism. We present two arguments against it, motivated from the interest in applying quantum probability (QP) theory in cognitive modeling. By QP theory, we imply the mathematics for assigning probabilities to events from quantum mechanics, without the physics. QP theory is a formal theory of probability, like CP theory. QP and CP theories are based on different axioms and so their predictions can diverge. QP theory is a plausible contender in decision making (and rationality). Recent work has shown that QP principles can provide the basis for simple, constrained models for empirical findings, which have been persistently problematic from a classical perspective, such as order effects on choice (Moore, 2002), the conjunction fallacy (Tversky and Kahneman, 1983), and the disjunction fallacy (Shafir and Tversky, 1992), for example, in Pothos and Busemeyer (2009), Trueblood and Busemeyer (2011), and Wang and Busemeyer (2013). Moreover, QP principles have been successfully

applied in other areas of cognition, such as memory (Bruza et al., 2009), perception (Atmanspacher and Filk, 2010), and conceptual combination [Aerts (2009); overviews in Busemeyer and Bruza (2011), Pothos and Busemeyer (2013), Wang et al. (2013)].

The above points attest to the descriptive status of QP theory in cognitive theory, not its normative status. Nevertheless, they motivate a consideration of explanatory concepts from QP theory in psychological debates. Of relevance presently is the idea of incompatibility, in relation to two (or more) questions (or possibilities etc.). According to classical theory, all questions are compatible, which means that the answer to any set of questions can be known concurrently. Following from Tversky and Kahneman's (1983) experiment which led to the finding of the conjunction fallacy, for example, classically, it can be established that Linda is a bank teller at the same time as deciding whether Linda is a feminist. As a result, it is always possible to specify a joint probability distribution for the outcomes of any arbitrary set of questions (this is the principle of unicity; Griffiths, 2003). The intuition that such questions are compatible appears obvious. How could it possibly be otherwise? Yet, in QP theory questions can be compatible or incompatible. In the latter case, certainty about one inexorably causes uncertainty about the other. Thus, resolving the question about whether Linda is a bank teller requires that we are uncertain about whether she is a feminist, and vice versa. Incompatibility means that there is no single sample space against which we can assess all possible questions about a system of interest (such as Linda). Rather, certainty about a particular question creates a novel perspective (sample space), against which the remaining questions can be assessed (these ideas broadly resonate with Evans's, 2006, 2007, "singularity" principle). Equally, incompatibility implies that it is impossible to define a joint probability distribution for the corresponding questions. One can only define a probability for a sequence of two events, which is order dependent.

The above leads us to our first point. We think that the representational requirements from the principle of unicity are cognitively unrealistic. If we imagine a representation space in which all question outcomes are compatible, then, for two questions, each axis corresponds to a particular combination of outcomes (one axis would correspond to the combination that Linda is a bank teller and not a feminist, etc.). For two questions with binary outcomes, we need a four dimensional space. The consideration of each additional binary question increases the dimensionality of the space by a factor of two, so that, for N binary questions, we require 2^N dimensions. A classical space for just 10 binary questions requires over 1000 dimensions. The CP theory requirements for representational capacity appear too stringent. Another way to look at this issue is that, regardless of the number of questions considered, classically it is always possible to construct a complete joint probability distribution. But, where would the information come from to construct such a joint probability distribution, especially when considering unfamiliar combinations of questions (such as being a bank teller or a feminist)? Note, the principle of indifference cannot provide a general solution to this issue (e.g., Gilboa, 2009).

Thus, we suggest that cognitively it is more plausible to consider some questions, especially ones not typically considered together, as incompatible. Indeed, there have been suggestions that, with practice, some of the decision making fallacies attenuate (Nilsson et al., 2014; Trueblood, pers. commu.). This conclusion reduces the plausibility that CP theory provides a good descriptive framework for decision making. By implication, QP theory is perhaps a framework for bounded rationality Simon, 1955: Perhaps not as rational as in principle possible (assuming CP theory is the ultimate standard of rationality), but the best that can be achieved, given (broadly assumed) limitations in the representational capacity of the cognitive system.

This discussion leads to our second point: exactly what is the evidence that probabilistic inference on the basis of CP theory is as *accurate as possible*? An a priori argument is the DBT. The consistency in probabilistic inference, which is demonstrated with the DBT, perhaps implies accuracy as well (i.e., do CP theory probabilities match empirical data?). Is it possible to prove a version of the DBT for QP theory as well? Superficially, this may appear not to be the case. First, the axioms of CP theory (on the basis of which the DBT is proved) are very different from those of QP theory. Second, verifiably (e.g., Gilio and Over, 2012), a classical decision maker, committing the conjunction fallacy in Tversky and Kahneman's (1983) experiment, is subject to a Dutch Book, that is, it is possible to specify a combination of stakes for the various hypotheses (Linda is a bank teller; Linda is a feminist; Linda is a bank teller and a feminist), which lead to a sure loss (or gain). However, it is possible to express the requirements for the DBT in terms of the fundamental principles of QP theory. Moreover, it is certainly true that if the questions about Linda are compatible (i.e., if we assume all events can be placed within the same sample space), then a Dutch Book is possible. But, if they are incompatible this is no longer necessarily the case, because the probabilities involved are based on different conditions (orders of evaluation). With work in progress, we are formalizing the relevant intuitions, but the idea is that accepting one incompatible outcome for Linda (e.g., that she is feminist) creates a separate sample space for another (e.g., that she is a bank teller).

We return to the question of the accuracy of probabilistic inference, since, ultimately, this must be the standard against which we assess whether CP theory or QP theory provide a better framework for understanding rationality. Our view is this: if all the relevant questions are compatible, then rationality is best understood in terms of CP theory (actually, the predictions between CP theory and QP theory with compatible questions would be identical; but, if all questions are compatible, why consider QP theory?). However, if some of the questions are incompatible, then QP theory will provide more accurate predictions for probabilistic inference. For example, if some questions are incompatible, then order effects may arise in conjunctions (Trueblood and Busemeyer, 2011; Wang and Busemeyer, 2013), while conjunction in CP theory is commutative (order effects can arise classically, but not without e.g., a conditionalization depending on order, which is unlikely to be known a priori). There are many effects of this kind, that is, ways in which the knowledge

that two questions are incompatible can lead us to probabilistic predictions divergent from those using CP theory. Thus, the question of whether QP theory is a better or worse standard for rational decision making, compared to CP theory, boils down to whether there are questions which are incompatible or not (cf. Oaksford, 2013). This is an exciting empirical issue.

ACKNOWLEDGMENTS

Emmanuel M. Pothos was supported by Leverhulme Trust grant RPG-2013-004 and Jerome R. Busemeyer by NSF grant ECCS—1002188. Emmanuel M. Pothos and Jerome R. Busemeyer were supported by Air Force Office of Scientific Research (AFOSR), Air Force Material Command, USAF, grants FA 8655-13-1-3044 and FA 9550-12-1-0397 respectively. The U.S Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright notation thereon.

REFERENCES

Aerts, D. (2009). Quantum structure in cognition. *J. Math. Psychol.* 53, 314–348. doi: 10.1016/j.jmp.2009.04.005

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychol. Rev.* 98, 409–429. doi: 10.1037/0033-295X.98.3.409

Atmanspacher, H., and Filk, T. (2010). A proposed test of temporal nonlocality in bistable perception. *J. Math. Psychol.* 54, 314–321. doi: 10.1016/j.jmp.2009.12.001

Bruza, P. D., Kitto, K., Nelson, D., and McEvoy, C. L. (2009). Is there something quantum-like about the human mental lexicon? *J. Math. Psychol.* 53, 362–377. doi: 10.1016/j.jmp.2009.04.004

Busemeyer, J. R., and Bruza, P. (2011). *Quantum Models of Cognition and Decision Making*. Cambridge: Cambridge University Press.

Chater, N., and Oaksford, M. (1993). Logicism, mental models and everyday reasoning: reply to garnham. *Mind Lang.* 8, 72–87. doi: 10.1111/j.1468-0017.1993.tb00271.x

Elqayam, S., and Evans, St. B. T. (2013). Rationality in the new paradigm: strict versus soft Bayesian approaches. *Think. Reason.* 19, 453–470. doi: 10.1080/13546783.2013.834268

Elqayam, S., and Evans, St. J. S. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of the human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Evans, St. J. B. T. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Hove: Psychology Press.

Evans, St. J. S. (2006). The heuristic-analytic theory of reasoning: extension and evaluation. *Psychon. Bull. Rev.* 13, 378–395. doi: 10.3758/BF03193858

Gilboa, I. (2009). *Theory of Decision Under Uncertainty*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511840203

Gilio, A., and Over, D. (2012). The psychology of inferring conditionals from disjunctions: a probabilistic study. *J. Math. Psychol.* 56, 118–131. doi: 10.1016/j.jmp.2012.02.006

Griffiths, R. B. (2003). *Consistent Quantum Theory*. Cambridge: Cambridge University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* 14, 357–364. doi: 10.1016/j.tics.2010.05.004

Howson, C., and Urbach, P. (1993). *Scientific Reasoning: The Bayesian Approach*. Chicago, IL: Open Court.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185

Moore, D. W. (2002). Measuring new types of question-order effects. *Public Opin. Q.* 66, 80–91. doi: 10.1086/338631

Nilsson, H., Rieskamp, J., and Jenny, M. A. (2014). Exploring the overestimating of conjunctive probabilities. Front. *Psychol.* 4:101. doi: 10.3389/fpsyg.2013.00101

Oaksford, M. (2013). Quantum probability, intuition, and human rationality. *Behav. Brain Sci.* 36, 303. doi: 10.1017/S0140525X12003081

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198524496.001.0001

Pothos, E. M., and Busemeyer, J. R. (2009). A quantum probability explanation for violations of "rational" decision theory. *Proc. Biol. Sci.* 276, 2171–2178. doi: 10.1098/rspb.2009.0121

Pothos, E. M., and Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behav. Brain Sci.* 36, 255–327. doi: 10.1017/S0140525X12001525

Shafir, E., and Tversky, A. (1992). Thinking through uncertainty: nonconsequential reasoning and choice. *Cogn. Psychol.* 24, 449–474. doi: 10.1016/0010-0285(92)90015-T

Simon, H. A. (1955). A behavioral model of rational choice. *Q. J. Econ.* 69, 99–118. doi: 10.2307/1884852

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 11, 3–22. doi: 10.1037/0033-2909.119.1.3

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Trueblood, J. S., and Busemeyer, J. R. (2011). A comparison of the belief-adjustment model and the quantum inference model as explanations of order effects in human inference. *Cogn. Sci.* 35, 1518–1552. doi: 10.1111/j.1551-6709.2011.01197.x

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjuctive fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Vineberg, S. (2011). "Dutch book arguments," in *The Stanford Encyclopedia of Philosophy Summer 2011 Edn*, ed Edward N. Zalta. Available online at: http://plato.stanford.edu/archives/sum2011/entries/dutch-book

Wakker, P. P. (2010). *Prospect Theory for Risk and Ambiguity*. Cambridge: Cambridge University Press.

Wang, Z., Busemeyer, J. R., Atmanspacher, H., and Pothos, E. M. (2013). The potential of using quantum theory to build models of cognition. *Top. Cogn. Sci.* 5, 672–688. doi: 10.1111/tops.12043

Wang, Z. A., and Busemeyer, J. R. (2013). Empirical tests of a quantum probability model for question order effects found in survey research. *Top. Cogn. Sci.* 5, 689–710. doi: 10.1111/tops.12040

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Q. J. Exp. Psychol.* 12, 129–140. doi: 10.1080/17470216008416717

Received: 07 November 2013; accepted: 15 January 2014; published online: 30 January 2014.

*Citation: Pothos EM and Busemeyer JR (2014) In search for a standard of rationality. Front. Psychol. **5**:49. doi: 10.3389/fpsyg.2014.00049*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

PERSPECTIVE ARTICLE
published: 07 October 2013
doi: 10.3389/fpsyg.2013.00710

Exploration, novelty, surprise, and free energy minimization

Philipp Schwartenbeck*, Thomas FitzGerald, Raymond J. Dolan and Karl Friston

The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Philipp Sterzer, University Hospital Charité, Germany
Shane Mueller, Michigan Technological University, USA
Mike Oaksford, University of London, UK

****Correspondence:***
Philipp Schwartenbeck, The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, 12 Queen Square, London, WC1N 3BG, UK
e-mail: philipp.schwartenbeck.12@ucl.ac.uk

This paper reviews recent developments under the free energy principle that introduce a normative perspective on classical economic (utilitarian) decision-making based on (active) Bayesian inference. It has been suggested that the free energy principle precludes novelty and complexity, because it assumes that biological systems—like ourselves—try to minimize the long-term average of surprise to maintain their homeostasis. However, recent formulations show that minimizing surprise leads naturally to concepts such as exploration and novelty bonuses. In this approach, agents infer a policy that minimizes surprise by minimizing the difference (or relative entropy) between likely and desired outcomes, which involves both pursuing the goal-state that has the highest expected utility (often termed "exploitation") and visiting a number of different goal-states ("exploration"). Crucially, the opportunity to visit new states increases the value of the current state. Casting decision-making problems within a variational framework, therefore, predicts that our behavior is governed by both the entropy and expected utility of future states. This dissolves any dialectic between minimizing surprise and exploration or novelty seeking.

Keywords: active inference, exploration, exploitation, novelty, reinforcement learning, free energy

INTRODUCTION

The free energy principle is a theoretical formulation of biological systems and their behavior (Friston et al., 2006; Friston, 2009, 2010) that has attracted much current research interest (Brown and Friston, 2012; Adams et al., 2013a; Apps and Tsakiris, 2013; Joffily and Coricelli, 2013; Moran et al., 2013). Its underlying premise is that a biological system, in order to underwrite its existence and avoid the dispersion of its physical states, has to maintain its states within certain bounds and, therefore, maintain a homeostasis. Under ergodic assumptions this means that it has to minimize its long-term average surprise (i.e., Shannon entropy) over the states it visits. Surprise is an information theoretic quantity that can be approximated with variational free energy (Feynman, 1972; Hinton and van Camp, 1993). Every system that maintains itself conforms to the imperative of minimizing the surprise associated with the states it encounters. In the context of neuroscience, this implies that the brain becomes a model of the world in order to evaluate surprise in relation to model-based predictions (Friston, 2012). Practically, this means that it has to elaborate internal predictions about sensory input and update them based on prediction errors, a process that can be formulated as generalized Bayesian filtering or predictive coding in the brain (Friston, 2005). The notion of active inference translates predictive coding into an embodied context and argues that surprise can be minimized in two ways: either by optimizing internal predictions about the world (perception) or via acting on the world to change sensory samples so that they match internal predictions (action) (Brown et al., 2011).

The premise that every biological system—such as the brain—has to minimize variational free energy promises to provide a unified account of brain function and behavior and has proven useful for understanding neuroanatomy, neurophysiology (Feldman and Friston, 2010; Bastos et al., 2012; Brown and Friston, 2012; Adams et al., 2013b; Moran et al., 2013), and psychiatry (Edwards et al., 2012; Adams et al., 2013a). However, many recent discussions have deconstructed and critiqued the theory (Clark, 2013). In particular, a recurring criticism runs as follows: if our main objective is to minimize surprise over the states and outcomes we encounter, how can this explain complex human behavior such as novelty seeking, exploration, and, furthermore, higher level aspirations such as art, music, poetry, or humor? Should we not, in accordance with the principle, prefer living in a highly predictable and un-stimulating environment where we could minimize our long-term surprise? Shouldn't we be aversive to novel stimuli? As it stands, this seems highly implausible; novel stimuli are sometimes aversive, but often quite the opposite. The challenge here is to reconcile the fundamental imperative that underlies self-organized behavior with the fact that we avoid monotonous environments and actively explore in order to seek novel and stimulating inputs (Kakade and Dayan, 2002).

The free energy principle—under which our theoretical arguments are developed—is the quintessential normative theory for action and perception. It is normative in the sense that it provides a well-defined objective function (variational free energy) that is optimized both by action and perception. Having said this, the normative aspect of free energy minimization (and implicit active inference) is complemented by a neuronally plausible implementation scheme, in the form of predictive coding. We do not focus on the underlying imperatives for minimizing free energy (this

has been fully addressed elsewhere). In this essay, we look specifically at the normative implications for behavior in the context of classical (economic) decision-making problems. Our normative account argues that optimal decisions minimize the relative entropy between likely and desired outcomes. This means that—in some contexts—agents are compelled to seek novel states, whereas in other contexts they maximize expected utility. We hope to show that explorative behavior is not just in accordance with the principle of free energy minimization but is in fact mandated when minimizing surprise (or maximizing model-evidence) in the context of decision-making behavior. In brief, we argue that when a policy (i.e., an action selection rule that entails a sequence of actions) is selected—in a way that includes uncertainty about outcomes—there is necessarily an exploratory drive that accompanies the classical maximization of expected utility.

BOREDOM AND NOVELTY SEEKING UNDER THE FREE ENERGY PRINCIPLE

When addressing this issue, one has to appreciate an important but subtle difference between two questions: one being why the imperative to minimize surprise does not predict that we seek out an impoverished or senseless environment; the other being how the free energy principle motivates the active exploration of new states. The former question is associated with the "dark room problem," which has been dealt with previously (Friston et al., 2009, 2012). The "dark room problem," however, does not refer to a real "problem" but merely a misapprehension about what is meant by "surprise": it can be easily resolved by appreciating the difference between minimizing the long-term surprise over states (i.e., the Shannon entropy) $H[S]$ *per se* compared to minimizing the long-term surprise given a specific (generative) model: $H[S|m]$. This means that agents are equipped with prior beliefs—which can be innate and acquired by natural selection such as an aversion to hypoglycaemia or dehydration or shaped by learning according to experience (Friston, 2011, 2013)—that define what an agent regards as surprising. Put simply, (most) agents would find it highly surprising to be incarcerated in a dark room and would thus generally try to avoid that state of affairs. More formally, there is a fundamental difference between the intuitive meaning of "surprise" in terms of unpredictable sensory input and surprise (in information theoretic terms) under a particular model of the world. Finding ourselves in a dark room (and being subject to a surprising sense of starvation and sensory deprivation) is a highly surprising state, even though it represents an environment with maximally predictable sensory input.

It is reassuring that the free energy principle does not compel us to seek an empty room, turn off the light and wait there until we die. However, what does it have to say concerning autonomous, purposeful behavior and why we actively aspire (to a certain extent) to novel, complex states? Why do we enjoy going to exhibitions and seeing our favorite piece of art—or learning about new artists—when our main objective is to restrict our existence to a limited number of (attractor) states to maintain a homeostasis?

This question is addressed in a recent application of the free energy framework, which casts complex, purposeful decision-making as active inference (Friston et al., 2013). The basic assumption that action minimizes surprise by selective sampling of sensory input (to match internal predictions) is applied to fictive states in the future. Put simply, this means that an agent's prior beliefs include the notion that it will act to minimize surprise. By analogy to perceptual inference—where agents are equipped with a generative model mapping from hidden causes to sensory consequences—the agent's generative model includes hidden (future) states and actions that the agent might perform (and their consequences). This implies that the agent has to represent itself in future states performing specific actions. In other words, it necessarily implies a model with a sense of agency.

Based on its generative model, the agent has to infer policies in order to minimize surprise about future outcomes. Beliefs about the (optimal) policies it will find itself pursuing is based on their value, which can be expressed in the following probabilistic terms:

$$Q(\pi|s_t) = -D_{KL}\left[P(s_T|s_t, \pi) \,\|P(s_T|m)\right] \quad (1)$$

Equation 1 formalizes the intuitive notion that valuable policies minimize the difference between likely and desired outcomes by bringing the former as close as possible to the latter. The left side of the equation refers to the value Q of a given policy π from a specific state at time $t \in T$. The right side of the equation defines this value as the (negative) difference or relative entropy (Kullback-Leibler Divergence) between two probability distributions: $P(s_T|s_t, \pi)$ refers to a probability distribution over outcome states, given a specific policy and a current state, whereas $P(s_T|m)$ refers to a probability distribution over outcomes based solely on the prior beliefs or intentional goals of the agent. The former distribution refers to the empirical prior over outcomes given a specific sequence of actions the agent might perform, whereas the latter refers to priors that represent which goal state the agent believes it will (desires to) attain. These goal priors are fixed and do not depend on sensory input: they represent a belief concerning states the agent will end up in. Desired goal states will be accorded a high probability (log-likelihood) of being encountered, resulting in a low surprise when this state is indeed visited. Undesirable states, by contrast, will be assigned with a low prior probability and therefore become highly surprising.

Crucially, casting decision-making as KL control (or equivalently surprise or relative-entropy minimization) subsumes classical notions of reward and utility that are central to fields such as behavioral economics and reinforcement learning, since the value of a state becomes simply a function of how surprising it is: visiting unsurprising states is associated with a high reward in classical reinforcement and utilitarian schemes, whereas surprising future states have low reward or high cost. In the framework of active inference, therefore, agents do not try to maximize reward but minimize surprise (about future states). Similar accounts following KL control have been proposed earlier (Solway and Botvinick, 2012; Huang and Rao, 2013), but in the context of prior beliefs over the magnitude of rewards in future states. Here, valuable policies minimize the Kullback-Leibler divergence (relative entropy) between the distribution of likely outcomes and the distribution of desired outcomes represented as belief about attaining them (which we assume to be fixed and defined a priori).

To see why this scheme mandates both exploitation (value maximization) and exploration (visiting novel states), one can rewrite this KL-Divergence term as:

$$
\begin{aligned}
-D_{KL}\left[P\left(s_T|s_t,\,\pi\right)\|P\left(s_T|m\right)\right] &= \sum_{s_T} P\left(s_T|s_t,\pi\right)\ln\frac{P\left(s_T|m\right)}{P(s_T|s_t,\pi)} \\
&= \sum_{s_T}(-P\left(s_T|s_t,\pi\right)\cdot\ln P\left(s_T|s_t,\pi\right) \\
&\quad + P\left(s_T|s_t,\pi\right)\cdot\ln P\left(s_T|m\right)) \\
&= -\sum_{s_T}(P\left(s_T|s_t,\pi\right)\cdot\ln P\left(s_T|s_t,\pi\right)) \\
&\quad + \sum_{s_T}(P\left(s_T|s_t,\pi\right)\cdot\ln P\left(s_T|m\right)) \\
&= H[P(s_T|s_t,\pi)] \\
&\quad + \sum_{s_T} P\left(s_T|s_t,\pi\right)\cdot u\left(s_T|m\right) \qquad (2)
\end{aligned}
$$

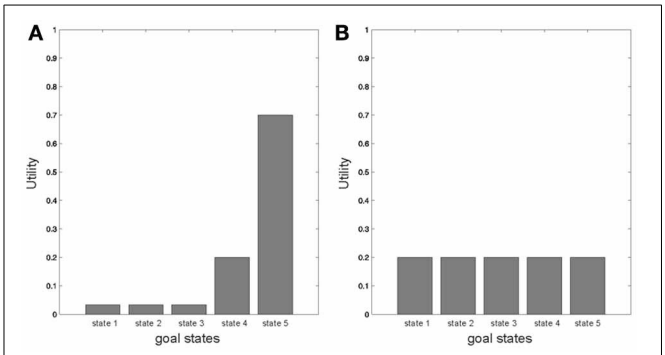


FIGURE 1 | Illustrating two different contexts in which a valuable policy is defined according to the likelihood of visiting a goal state with highest utility (A–because one goal state is clearly favored by the agent) or according to the likelihood of visiting many different states (B–because no goal state is favored), depending on the different representations of utilities of goal states in an agent's prior beliefs.

This decomposition of the value of a policy is important as it speaks to two different ways of maximizing the value of a selected policy: the first term is the entropy over goal-states, which reflects the number of different outcomes the agent is likely to experience under a specific policy, whereas the second term represents the expected utility over outcomes that depends on an agent's priors $u\left(s_T|m\right) = \ln P\left(s_T|m\right)$, which constitute an agent's goals and the (beliefs about) utility of final states. This term increases the value of a policy that secures the outcome with highest expected utility. The relative contribution of these two to the value of a policy depends on the current state and the precision with which prior beliefs about goals are held, as illustrated in **Figure 1**. When the utilities of outcomes differ (and are well-defined), a policy that makes visiting the outcome with highest utility (and only this one) most likely will be the most valuable. When outcomes have the same or similar utilities, on the other hand, policies cannot be differentiated according to expected utility. In this case, policies will be valuable if they maximize the entropy over outcome states in accordance with the maximum entropy principle (Jaynes, 1957), which means the agent will try to visit all states with equal probability.

The notion behind this decomposition goes beyond stating that agents maximize entropy if the utilities of outcomes are the same and maximize expected utility if they differ, but rather implies that all our decisions are influenced by entropy and expected utility—with a context-sensitive weighting of those two. This decomposition may account for numerous instances of every-day choice behavior, such as why we appreciate variation over outcomes much more when we buy a chocolate bar as opposed to a car: when the differences in the expected utilities of outcomes become less differentiable, agents will try to visit several states and not just the state that has highest utility.

This distinction is interesting because it maps to various other accounts of complex decision-making and planning. Most importantly, this distinction resembles exploration-exploitation (Sutton and Barto, 1998; Cohen et al., 2007), which is prominent in reinforcement learning paradigms. Here, choosing a policy that maximizes expected utility corresponds to exploitation, whereas maximizing entropy over outcomes corresponds to exploration. An important difference is, however, that exploration is often equated with random or stochastic behavior in reinforcement learning schemes (but see Thrun, 1992), whereas in our framework, maximizing entropy over outcome states is a goal-driven, purposeful process—with the aim of accessing allowable states. Furthermore, this distinction neatly reflects the differentiation between intrinsic and extrinsic reward (Schmidhuber, 1991, 2009; Luciw et al., 2013), where extrinsic reward refers to externally administered reinforcement—corresponding to maximizing expected utility—and intrinsic reward is associated with maximizing entropy over outcomes. Maximizing intrinsic reward is usually associated with seeking new experiences in order to increase context-sensitive learning—which is reflected as increasing model-evidence or minimizing surprise in the active inference framework.

The formal difference between classical (utilitarian) formulations of valuable behavior and those that are deemed valuable under active inference can be reduced to a simple distinction: in classical schemes, policies are chosen to maximize expected utility, whereas in active inference they are chosen to minimize the probabilistic divergence between controlled outcomes and a probability distribution that is defined in terms of utility. This difference induces an entropy or exploration term that would require some *ad-hoc* augmentation of classical utility functions. However, there is something more fundamental about the different approaches. Recall from above that policies are inferred during active inference. In other words, the agent has to infer which policy it is most likely to pursue and then selects that policy. Because this formulation converts an optimal control or reinforcement learning problem into an inference problem, beliefs about optimal policies can themselves be optimized in terms of their precision or confidence. This precision corresponds to the temperature or sensitivity parameter in classical models that appeal to softmax choice rules. This is important because precision can be optimized in a Bayes optimal sense during active

inference and ceases to be an *ad-hoc* or descriptive parameter of choice behavior. In Friston et al. (2013) we show that the updating of precision has many of the hallmarks of dopamine discharges.

CONCLUSION

The aim of this paper was to explain exploration and novelty seeking under the free energy principle. The formalism presented here is part of a general framework of decision making as active inference and will be discussed in more detail elsewhere (Friston et al., 2013). This theoretical piece serves to underlie the basic issues and potential ways forward. It will be complemented by a series of more technical papers (based on simulations, empirical studies of choice behavior and functional neuroimaging) that provide specific examples and operationalize the ideas discussed in the current overview. We have shown that concepts like intrinsic and extrinsic reward—or exploration and exploitation—emerge naturally from casting decision-making under the normative assumption that agents minimize the relative entropy (KL-divergence) between likely and desired outcomes. Valuable policies will maximize expected utility or entropy over outcomes (or both), where the relative weight of these two mechanisms is context specific and depends upon prior beliefs.

We therefore resolve an apparent paradox concerning the incompatibility of minimizing surprise and the exploration of novel states, which constitute an essential aspect of human and animal behavior. Indeed, under certain circumstances, surprise can be minimized (i.e., model evidence can be maximized) if an agent selects a policy that increases the likelihood of visiting new and informative states. The concept of surprise minimization, therefore, by no means precludes agents from active exploration or appreciating novelty but rather explicitly predicts that this is an important factor in guiding our behavior. The most straightforward application of the formalism presented here clearly lies in economic decision-making tasks. Our formalism is certainly not sufficient–in the given form–to explain all aspects of higher level activities, such as the appreciation of fine arts. Maximizing intrinsic reward and visiting new and informative states to maximize model evidence (i.e., improve our model of the world) may, however, lay the foundation for future developments along these lines. Furthermore, empirical research is currently investigating the relative influence of entropy and expected utility maximization on behavior and their association with neuronal activation—which may well be related to specific personality traits such as sensation seeking. We look forward to reporting these results in the not too distant future.

ACKNOWLEDGMENTS

We would like to thank Francesco Rigoli and Timothée Devaux for helpful discussions and insightful comments on this matter. Furthermore, we would like to thank the reviewers of this manuscript for their detailed and helpful suggestions on the presentation of these ideas.

This work was supported by the Wellcome Trust (Raymond J. Dolan Senior Investigator Award 098362/Z/12/Z). The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust 091593/Z/10/Z.

REFERENCES

Adams, R., Shipp, S., and Friston, K. J. (2013a). Predictions not commands: active inference in the motor system. *Brain Struct. Funct.* 218, 611–643. doi: 10.1007/s00429-012-0475-5

Adams, R., Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013b). The computational anatomy of psychosis. *Front. Psychiatry* 4:47. doi: 10.3389/fpsyt.2013.00047

Apps, M. A. J., and Tsakiris, M. (2013). The free-energy self: a predictive coding account of self-recognition. *Neurosci. Biobehav. Rev.* doi: 10.1016/j.neubiorev.2013.01.029. [Epub ahead of print].

Bastos, A. M., Usrey, W. M., Adams, R., Mangun, G. R., Fries, P., and Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Brown, H., Friston, K., and Bestmann, S. (2011). Active inference, attention, and motor preparation. *Front. Psychol.* 2:218. doi: 10.3389/fpsyg.2011.00218

Brown, H., and Friston, K. J. (2012). Free-energy and illusions: the cornsweet effect. *Front. Psychol.* 3:43. doi: 10.3389/fpsyg.2012.00043

Clark, A. (2013). Whatever next. Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Cohen, J. D., McClure, S. M., and Yu, A. J. (2007). Should I stay or should I go. How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 362, 933–942. doi: 10.1098/rstb.2007.2098

Edwards, M., Adams, R., Brown, H., Pareés, I., and Friston, K. (2012). A Bayesian account of "hysteria." *Brain* 135, 3495–3512. doi: 10.1093/brain/aws129

Feldman, H., and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* 4:215. doi: 10.3389/fnhum.2010.00215

Feynman, R. A. (1972). *Statistical Mechanics.* Reading, MA: Benjamin.

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K. (2009). The free-energy principle: a rough guide to the brain. *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory. *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2011). "Embodied Inference: or 'I think therefore I am, if I am what I think,'" in *The Implications of Embodiment (Cognition and Communication),* eds W. Tschacher and C. Bergomi (Exeter: Imprint Academic), 89–125.

Friston, K. (2012). A free energy principle for biological systems. *Entropy* 14, 2100–2121. doi: 10.3390/e14112100

Friston, K. (2013). Active inference and free energy. *Behav. Brain Sci.* 36, 212–213. doi: 10.1017/S0140525X12002142

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol.* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Front. Psychol.* 3:130. doi: 10.3389/fpsyg.2012.00130

Friston. K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* 7:598. doi: 10.3389/fnhum.2013.00598

Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference. *PLoS ONE* 4:e6421. doi: 10.1371/journal.pone.0006421

Hinton, G. E., and van Camp, D. (1993). "Keeping the neural networks simple by minimizing the description length of the weights," in *Proceedings of the Sixth Annual Conference on Computational Learning Theory–COLT'93* (New York, NY: ACM Press), 5–13. doi: 10.1145/168304.168306

Huang, Y., and Rao, R. P. N. (2013). Reward optimization in the primate brain: a probabilistic model of decision making under uncertainty. *PLoS ONE* 8:e53344. doi: 10.1371/journal.pone.0053344

Jaynes, E. (1957). Information theory and statistical mechanics. *Phys. Rev.* 106, 620–630. doi: 10.1103/PhysRev.106.620

Joffily, M., and Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comput. Biol.* 9:e1003094. doi: 10.1371/journal.pcbi.1003094

Kakade, S., and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Netw.* 15, 549–559. doi: 10.1016/S0893-6080(02)00048-5

Luciw, M., Kompella, V., Kazerounian, S., and Schmidhuber, J. (2013). An intrinsic value system for developing multiple invariant representations with incremental slowness learning. *Front. Neurorobot.* 7:9. doi: 10.3389/fnbot.2013.00009

Moran, R. J., Campo, P., Symmonds, M., Stephan, K. E., Dolan, R. J., and Friston, K. (2013). Free energy, precision and learning: the role of cholinergic neuromodulation. *J. Neurosci.* 33, 8227–8236. doi: 10.1523/JNEUROSCI.4255-12.2013

Schmidhuber, J. (1991). "Curious model-building control systems," in *Proceedings of IEEE International Joint Conference on Neural Networks* (Singapore), 1458–1463.

Schmidhuber, J. (2009). Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. 計測と制御 48, 21–32.

Solway, A., and Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychol. Rev.* 119, 120–154. doi: 10.1037/a0026435

Sutton, R., and Barto, A. (1998). *Reinforcement Learning. An Introduction.* Cambridge, MA: MIT Press.

Thrun, S. B. (1992). *The Role of Exploration in Learning Control. Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*. Frorence, KY: Van Nostrand Reinhold.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 24 July 2013; paper pending published: 22 August 2013; accepted: 17 September 2013; published online: 07 October 2013.

*Citation: Schwartenbeck P, FitzGerald T, Dolan RJ and Friston K (2013) Exploration, novelty, surprise, and free energy minimization. Front. Psychol. **4**:710. doi: 10.3389/fpsyg.2013.00710*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Descriptivist perspectives

frontiers in
PSYCHOLOGY

PERSPECTIVE ARTICLE
published: 12 February 2014
doi: 10.3389/fpsyg.2014.00104

Rationality and the illusion of choice

Jonathan St. B. T. Evans*

School of Psychology, University of Plymouth, Plymouth, UK

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
David E. Over, Durham University, UK
Linden John Ball, University of Central Lancashire, UK

****Correspondence:***
Jonathan St. B. T. Evans, School of Psychology, University of Plymouth, Plymouth PL4 8AA, UK
e-mail: j.evans@plymouth.ac.uk

The psychology of reasoning and decision making (RDM) shares the methodology of cognitive psychology in that researchers assume that participants are doing their best to solve the problems according to the instruction. Unlike other cognitive researchers, however, they often view erroneous answers evidence of irrationality rather than limited efficiency in the cognitive systems studied. Philosophers and psychologists also talk of people being irrational in a special sense that does not apply to other animals, who are seen as having no choice in their own behavior. I argue here that (a) RDM is no different from other fields of cognitive psychology and should be subject to the same kind of scientific inferences, and (b) the special human sense of irrationality derives from folk psychology and the illusory belief that there are conscious people in charge of their minds and decisions.

Keywords: rationality, decision making, folk psychology, illusion of control, reasoning

INTRODUCTION

Two fields stand out as different within cognitive psychology. These are the study of reasoning, especially deductive reasoning and statistical inference, and the more broadly defined field of decision making. For simplicity I label these topics as the study of reasoning and decision making (RDM). What make RDM different from all other fields of cognitive psychology is that psychologists constantly argued with each other and with philosophers about whether the behavior of their participants is *rational* (see Cohen, 1981; Stanovich and West, 2000; Elqayam and Evans, 2011). The question I address here is why? What is so different about RDM that it attracts the interests of philosophers and compulsively engages experimental psychologists in judgments of how good or bad is the RDM they observe.

Let us first consider the nature of cognitive psychology in general. It is branch of cognitive science, concerned with the empirical and theoretical study of cognitive processes in humans. It covers a wide collection of processes connected with perception, attention, memory, language, and thinking. However, only in the RDM subset of the psychology of thinking is rationality an issue. For sure, *accuracy* measures are used throughout cognitive psychology. We can measure whether participants detect faint signals, make accurate judgments of distances, recall words read to them correctly and so on. The study of non-veridical functions is also a part of wider cognitive psychology, for example the study of visual illusions, memory lapses, and cognitive failures in normal people as well as various pathological conditions linked to brain damage, such as aphasia. But in none of these cases are inaccurate responses regarded as *irrational*. Visual illusions are attributed to normally adaptive cognitive mechanisms that can be tricked under special circumstances; memory errors reflect limited capacity systems and pathological cognition to brain damage or clinical disorders. In no case is the *person* held responsible and denounced as irrational[1].

Even in the psychology of thinking, the same approach prevails in many topic areas. For example, when we give people longer letter strings they increasing fail to find anagrams. We do not say that failing to solve a long anagram problem is irrational; indeed it would seem quite anomalous to do so. In fact, in the broader field of problem solving generally, despite obvious similarities with RDM, there is much measurement of error but no debate about rationality. We measure performance errors to investigate psychology mechanisms and their design limitations but not to declare people irrational as result. But if the psychology of problem solving needs no rationality debate, why is it that the study of RDM does?

NORM-REFERENCING IN COGNITIVE PSYCHOLOGY

A clear correlate of rationality debating within cognitive psychology is the prevalence of *norm-referencing*. In most of cognitive psychology there is little or no debate about what constitutes an error. A signal is present or not and hence detected or not by the participants' judgment; a word recalled was either present or absent in the list of words presented to the participant; an anagram offered either uses the letters presented or it does not. But the study of RDM is different in this respect. In these fields, experimenters need to apply a *normative theory* in order to decide whether an error has been made. If we divide cognitive psychology into fields that are norm-referenced and those that are not, there is an almost perfect correlation with the presence of rationality judgments.

It is important to note that normative theories are not psychological theories and that they derive from disciplines outside of psychology. For example, the dominant theory of rational decision making was derived from the disciplines of economics and mathematics (von Neumann and Morgenstern, 1944) and first introduced to psychologists by Edwards (1954). Study of decisions made under uncertainty, and the assessment of risk

[1]I am not saying that judgmental terms are entirely absent in other fields of psychology, for example with regard to false memories and unfounded beliefs. However, I believe that reasoning and decision making are the only topics in which rationality is a central concern.

became a mainstream topic for psychologists who attempted to assess conformity to rational principles, as defined by economists and mathematicians. A spin-off from this was to study people's intuitive grasp of statistical principles derived from the probability calculus, such as Bayes' theorem. While early assessment of people's intuitive statistical abilities were optimistic (Peterson and Beach, 1967), this soon changed when Tversky and Kahneman (1974) launched their heuristics and biases program in early 1970s (for later reports, see Kahneman et al., 1982; Gilovich et al., 2002).

Wason (1960, 1966, 1968) and Wason and Johnson-Laird (1972) famously attributed irrationality to his participants based on their frequent failure to solve his 2-4-6 and selection task problems (see Evans, 2002, for quoted examples). He described a verification bias, more generally known as confirmation bias, which he suggested was irrational as it failed to comply with Popper's strictures for good scientific thinking. None of this has stood the test of time as his verification bias account has been discredited for both tasks (see Evans, 2007a) and Popper's philosophy of science has been strongly challenged by Bayesian critics (Poletiek, 2001; Howson and Urbach, 2006). In a sense, however, that is beside the point. People were considered irrational because they appeared to violate a popular normative theory of the time (Popper, 1959). Similarly, studies of deductive reasoning from the 1980s onward have shown people to be illogical (Evans, 2007a; Manktelow, 2012) but again the use of standard logic has been challenged (e.g., Oaksford and Chater, 2007).

It is evident that the need to apply a normative theory creates problems that are not present in other parts of cognitive psychology because we can debate whether such theories are correctly formulated or appropriately applied. However, it is far from obvious to me why in itself this should lead to a rationality debate. Why is a person wrongly identifying a face merely mistaken, while a person failing to maximize utility or making a logical error irrational? As we have seen, in most parts of cognitive psychology, evidence of error is not seen as evidence of irrationality. In fact, it seems quite ludicrous to suggest, for example, that someone falling prey to a standard visual illusion is being irrational. So there must be more to this problem than simply the ambiguity involved with norm referencing.

RATIONALITY AND VOLITION

A pigeon that learns to peck at a key in order to obtain food pellets can be described as *instrumentally rational,* that is, acting in such a way as to achieve its goals. Instrumental rationality is also known sometimes as personal or individual rationality (Stanovich, 1999). In fact, the argument can be made that animals are *more* instrumentally rational than humans as defined by performance on judgment and decision making tasks (Stanovich, 2013). Humans, with their complex layers of multiple goals and value systems will not always choose correctly according to the immediate goals that the psychologists uses to determine rationality. Of course, we could argue that this is due more to incorrect applications of norm-referencing than superior rationality of animals.

If we consider animals a little more, it becomes clear that there is a curious lack of complementarity between the terms rational and irrational. Animals frequently follow instinctive behavior patterns which conflict with their individual interests, exposing themselves to injury or death in pursuit of the interests of their selfish genes. More accurately, they follow instructions which helped genes to replicate in their environment of evolutionary adaptation at some time in the past. So are animals behaving *irrationally* when they act (by genetic compulsion) in ways that violate their interests as individuals? Surely not, as they have no choice in the matter. As Stanovich (2011), p.3) puts it: "an animal can be arational, but only humans can be irrational." But if they are not irrational when they act against their interests, in what sense are they rational when they act for them? There is some sense of rationality, applicable to humans, which seems not to apply to non-human animals.

It seems to me that in this important and distinctly human sense of the term, rationality is not simply to do with instrumentality; it is to do with *choice*. I have written elsewhere on the theory that humans have an old mind, animal like in many ways, combined with a new and distinctively human mind (Evans, 2010, in press; see also R'eber, 1993; Epstein, 1994; Evans and Over, 1996; Stanovich, 2004 for examples of many related earlier works along these lines). The rationality of the old mind is very much like the rationality of animals. We, like them, learn habits and procedures from experience that enable us to repeat behaviors rewarded in the past. This provides us and them with a form of instrumental rationality. But new mind rationality is not the slave of the past; as humans we can imagine the futures, conduct thought experiments and mental simulations and choose to act in one way rather than another. We can also (sometimes) manage to override our old minds, inhibiting our wishes to smoke cigarettes, join gambling games and other activities which may feel quite compulsive but conflict the goals that are new mind is setting for our futures. In fact, we are most likely to praise someone as rational when the new mind overrides in this way and conversely quick to condemn as irrational, the people who give way to their basic urges. However, while new mind cognition is volitional that does not mean that the individual is free to choose actions in all circumstances. Our behavior is the product of both old and new minds and so powerful emotions and strong habits may override the choices of the new mind. It is also a mistake to equate the new mind with the conscious person (see Evans, 2010, Chap. 7).

Another issue here lies with the general methodology of cognitive psychology. All cognitive experiments study *intendedly rational behavior*. It is nothing distinctive to RDM that participants are assumed to understand the instructions and be attempting to comply with them. If they were not bothering, then we could not, for example, infer that failure to recall a word reflected a limitation in memory capacity. What is distinctive to RDM is that when people fail to find the correct answer (according to some normative theory) they are often deemed to be irrational. But the method *presupposes* new mind rationality (compliance with instructions, making best effort). How can we both presuppose rationality and then infer irrationality from errors? Researchers in no other fields of cognitive psychology do this, inferring instead cognitive limitations from errors.

There is nothing inherently different about RDM tasks that justifies this difference. If the assumption of intendedly rational behavior is sound for the study of lexical decisions, semantic

memory and size constancy, then it is also sound for the study of deductive reasoning, probability judgment and decision making. If RDM researchers can say that people did not really understand the instructions or were not doing their best of comply with them, then why should we assume that they were compliant in studies of the serial position curve? If – as seems much more likely – RDM researchers endorse the cognitive method and share its assumptions, then on what basis can they equate errors with irrationality? Is it the underlying cognitive mechanisms that cause irrational choices, despite the best efforts of the conscious person? But in what sense can a mechanism be said to be irrational? It can be well or badly designed, fit for purpose or not but surely it cannot have rationality.

Stanovich (2011, p. 5) is admirably clear on this point: "... rationality is a personal entity and not a subpersonal one ... A memory system in the human brain is not rational or irrational, it is merely efficient or inefficient,". So it would seem that rationality, in this special human sense, is a property of the *person*. But who or what exactly is the person? It is clearly not be equated with organism as a whole, nor with the brain. So my brain cannot be irrational, and nor can the mind defined as the whole working of the brain in terms of its cognitive processes. In my detailed account of the two minds theory, I describe the person as a construction of the new mind and in many ways an illusory one. The conscious person whom we feel ourselves to be is subject to illusion of control and intentions that have been cleverly demonstrated by researchers in social psychology (see Evans, 2010, Chap. 7).

FOLK PSYCHOLOGY AND TWO MINDS CONFLICT

I think it is time for me to propose an answer to the puzzle. What is it about RDM that provokes a rationality debate absent in the rest of cognitive psychology? I believe the answer lies in folk psychology, in the ingrained beliefs that we all hold about the human mind and its operation[2]. Folk psychology embodies what I call the Chief Executive Model of the mind (Evans, 2010). We think of ourselves and others as conscious people in charge of our decisions[3]. To be sure there are many automated and unconscious mechanisms responsible for such matters as language processing, pattern recognition, memory retrieval etc. But these are merely slave systems doing our bidding. We, the conscious persons, are still in charge, still calling the shots. This is a powerful illusion, but an illusion nonetheless. There is now much accumulated evidence that we lack knowledge of our mental processes and the reasons underlying our decisions, frequently rationalizing or theorizing about our own behavior (Wilson et al., 1993; Wilson, 2002). The feeling that we are in control and that conscious thought determines actions is also an illusion (Bargh and Ferguson, 2000; Velmans, 2000; Wegner, 2002).

In two minds theory (Stanovich, 2004; Evans, 2010) conflict can easily arise between the goals that are pursued in the new and old minds. Moreover, the cognitive mechanisms for pursuit of goals differ radically, with experiential learning dominating the old mind, and hypothetical thinking the new mind. Two minds conflict is the essential cause of the cognitive biases that are observed in the study of reasoning and decision making. Biases arise from automated and unconscious mechanisms which divert us from solution of the tasks set. Frequently, there is a default intuitive response that leads people into error unless overridden by conscious reasoning (Kahneman and Frederick, 2002; Frederick, 2005; Evans, 2007b; Stanovich, 2011; Thompson et al., 2011). The ability to override such defaults is influenced by a number of factors including confidence in the original answer, cognitive ability and thinking dispositions. But in general, when someone fails to reason correctly according to the instruction it is due to an unconscious or intuitive influence of some kind. They are not choosing to get the answer wrong[4].

Outside of the laboratory, the behavior that strikes us as irrational is that in which a person experiences a two minds conflict in which the old mind is winning. For example, the heavily obese, compulsive gamblers and alcoholics are treated with very little sympathy in modern society. They are held to be responsible for their own health or financial problems because they could apparently *choose* to be different. Those of us who are not problem gamblers, for example, think it quite *irrational* that people should continue to bet money on casino games like roulette. The normative theory agrees, because all betting systems are based on the fallacious belief that later bets can compensate for earlier ones, whereas each individual bet has an expected loss (Wagenaar, 1988). But from a psychological point of view this normative analysis is not only simplistic but essentially useless in understanding the causes of problem gambling and how to deal with them. Most effective in such cases is cognitive-behavioral therapy which is essentially a two minds treatment (see Evans, 2010, Chap. 8).

CONCLUSIONS

There is nothing wrong with normative theories in themselves, nor with the tendency to debate which one is appropriate for a particular task. It is useful to have a measurement of error in RDM for the same reason as in other fields of cognitive psychology. If our decisions are suboptimal, for example, we can ask what limitations of our cognitive mechanisms are responsible. Is it a capacity limitation, or lack of experience or relevant learning? I have no problem, for example, agreeing that neglecting base rates in Bayesian inference is an error (Barbey and Sloman, 2007). I do have great difficulty in seeing it as evidence for irrationality, however. If people have not studied statistics, do not know the equation of Bayes' theorem and are not able to do complicated calculations in their heads, it is not surprising they make errors. But why is this irrational? As Elqayam and Evans (2011) point out, it as though *learning* has been excluded from the equation. We must apparently be able to reason well without relevant training and learning in order to be judged as rational.

[2]Note that I am not restricting the use of the term "folk psychology" to belief-desire psychology as is common in the philosophical literature.

[3]Folk psychology is close to the (largely discredited) interactive dualism of Descartes on this point. If I am right about this, then he was essentially formalizing intuitions about conscious minds that we all share.

[4]Stanovich's (2011) analysis implies that the choice lies within the "rational" thinking dispositions of what he calls the reflective mind. My view is that such dispositions are personality characteristics that are not chosen by the "person". The fact that, as he correctly claims, such dispositions can be modified by education and training is neither here nor there.

The problem lies not in the use of normative systems as such but in equation of conforming to them as an indicator of rational thought. Perhaps this practice is inherited from disciplines like philosophy and economics from which our normative theories derive. But to me it does not justify the treatment of RDM as different from any other field of cognitive psychology. We are still studying intendedly rational behavior and if people make errors it is not because they could have chosen to do otherwise. The belief that people can be irrational in a special sense that does not apply to other animals derives, I believe, from an illusion in folk psychology that there are somehow conscious persons, distinct from their minds and brains, who are in control of their behavior. People are certainly in possession of minds that are limited, inefficient and not always well adapted to the task at hand. So they are not invariably rational in the way that Panglossian authors (e.g., Cohen, 1981) claim, meaning that people are invariably well adapted and optimized. But nor can people be *irrational* either, in the sense derived from folk psychology.

REFERENCES

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological validity to dual processes. *Behav. Brain Sci.* 30, 241–297.

Bargh, J. A., and Ferguson, M. J. (2000). Beyond behaviorism: on the automaticity of higher mental processes. *Psychol. Bull.* 126, 925–945. doi: 10.1037/0033-2909.126.6.925

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behav. Brain Sci.* 4, 317–370. doi: 10.1017/S0140525X00009092

Edwards, W. (1954). The theory of decision making. *Psychol. Bull.* 41, 380–417. doi: 10.1037/h0053870

Elqayam, S., and Evans, J. St. B. T. (2011). Subtracting "ought from 'is:" descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–290. doi: 10.1017/S0140525X1100001X

Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *Am. Psychol.* 49, 709–724. doi: 10.1037/0003-066X.49.8.709

Evans, J. St. B. T. (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychol. Bull.* 128, 978–996. doi: 10.1037/0033-2909.128.6.978

Evans, J. St. B. T. (2007a). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Hove: Psychology Press.

Evans, J. St. B. T. (2007b). On the resolution of conflict in dual-process theories of reasoning. *Think. Reason.* 13, 321–329. doi: 10.1080/13546780601008825

Evans, J. St. B. T. (2010). *Thinking Twice: Two Minds in One Brain*. Oxford: Oxford University Press.

Evans, J. St. B. T. (in press). Two minds rationality. Thinking & Reasoning. doi: 10.1080/13546783.2013.845605

Evans, J. St. B. T., and Over, D. E. (1996). *Rationality and Reasoning*. Hove: Psychology Press.

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19 , 25–42. doi: 10.1257/089533005775196732

Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511808098

Howson, C., and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*, 3rd Edn. Chicago: Open Court.

Kahneman, D., and Frederick, S. (2002). "Representativeness revisited: attribute substitution in intuitive judgement," in *Heuristics and Biases: The Psychology of Intuitive Judgment,* eds T. Gilovich, D. Griffin, and D. Kahneman (Cambridge: Cambridge University Press), 49–81. doi: 10.1017/CBO9780511808098.004

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511809477

Manktelow, K. I. (2012). *Thinking and Reasoning*. Hove, UK: Psychology Press.

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198524496.001.0001

Peterson, C. R., and Beach, L. R. (1967). Man as an intuitive statistician. *Psychol. Bull.* 68, 29–46. doi: 10.1037/h0024722

Poletiek, F. (2001). *Hypothesis-Testing Behaviour*. Hove, UK: Psychology Press.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

R'eber, A. S. (1993). *Implicit Learning and Tacit Knowledge*. Oxford: Oxford University Press.

Stanovich, K. E. (1999). *Who is Rational? Studies of Individual Differences in Reasoning*. Mahway, NJ: Lawrence Elrbaum Associates.

Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning the Age of Darwin*. Chicago: University of Chicago Press. doi: 10.7208/chicago/9780226771199.001.0001

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. New York: Oxford University Press.

Stanovich, K. E. (2013). Why humans are (sometimes) less rational than other animals: cognitive complexity and the axioms of rational choice. *Think. Reason.* 19, 1–26. doi: 10.1080/13546783.2012.713178

Stanovich, K. E., and West, R. F. (2000). Advancing the rationality debate. *Behav.Brain Sci.* 23, 701–726. doi: 10.1017/S0140525X00623439

Thompson, V. A., Prowse Turner, J. A., and Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognit. Psychol.* 63, 107–140. doi: 10.1016/j.cogpsych.2011.06.001

Tversky, A., and Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Velmans, M. (2000). *Understanding Consciousness*. London: Routledge. doi: 10.4324/9780203465028

von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Wagenaar, W. A. (1988). *Pardoxes of Gambling Behaviour*. Hove and London: Erlbaum.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Q. J. Exp. Psychol.* 12,129–140. doi: 10.1080/17470216008416717

Wason, P. C. (1966). "Reasoning," in *New Horizons in Psychology I*, ed. B. M. Foss (Harmandsworth: Penguin), 106–137.

Wason, P. C. (1968). "On the failure to eliminate hypotheses: a second look," in *Thinking and Reasoning,* eds P. C.Wason and P. N. Johnson-Laird (Harmandsworth: Penguin), 165–174.

Wason, P. C., and Johnson-Laird, P. N. (1972). *Psychology of Reasoning: Structure and Content*. London: Batsford.

Wegner, D. M. (2002). *The Illusion of Conscious Will*. Cambridge: MIT books.

Wilson, T. D. (2002). *Strangers to Ourselves*. Cambridge: Belknap Press.

Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J., and Lafleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Pers. Soc. Psychol. Bull.* 19, 331–339. doi: 10.1177/0146167293193010

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 November 2013; paper pending published: 03 January 2014; accepted: 26 January 2014; published online: 12 February 2014.

*Citation: Evans JSBT (2014) Rationality and the illusion of choice. Front. Psychol. **5**:104. doi: 10.3389/fpsyg.2014.00104*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

PERSPECTIVE ARTICLE
published: 22 July 2014
doi: 10.3389/fpsyg.2014.00799

How (not) to draw philosophical implications from the cognitive nature of concepts: the case of intentionality

***Kazuki Iijima*[1,2] * *and Koji Ota*[3]**

[1] *Brain Science Institute, Tamagawa University, Tokyo, Japan*
[2] *Japan Society for the Promotion of Science, Tokyo, Japan*
[3] *Department of Basic Science, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan*

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Maria Olkkonen, University of Pennsylvania, USA
Joshua Knobe, Yale University, USA

****Correspondence:***
Kazuki Iijima, Brain Science Institute, Tamagawa University, 6-1-1 Tamagawa-Gakuen, Machida, Tokyo 194-8610, Japan
e-mail: iijima.kazuki@14.alumni.u-tokyo.ac.jp

Philosophers have often appealed to intuitive judgments in various thought experiments to support or reject particular theses. Experimental philosophy is an emerging discipline that examines the cognitive nature of such intuitive judgments. In this paper, we assess the methodological and epistemological status of experimental philosophy. We focus on the Knobe effect, in which our intuitive judgment of the intentionality of an action seems to depend on the perceived moral status of that action. The debate on the philosophical implications of the Knobe effect has been framed in terms of the distinction between the competence and performance of the concept of intentionality. Some scholars seem to suggest that the Knobe effect reflects the competence (or otherwise, the performance error) of the concept of intentionality. However, we argue that these notions are purely functional and thus do not have philosophical implications, without assuming normativism, which we see as problematic in a psychological methodology. Finally, focusing on the gap between competence and rationality, we suggest future directions for experimental philosophy.

Keywords: experimental philosophy, normativism, descriptivism, Knobe effect, intentionality, theory of mind

COMPETENCE AND PERFORMANCE IN EXPERIMENTAL PHILOSOPHY

Since the beginning of the 21st century, a new research program, "experimental philosophy," which systematically studies the nature of intuition with psychological methodologies, has become increasingly popular (Knobe and Nichols, 2008a; Alexander, 2012; Knobe et al., 2012; Knobe and Nichols, 2013). In this paper, we consider why philosophical arguments are not informed by the identification of the cognitive natures of philosophical concepts. In what follows, while we mainly focus on the concept of intentionality, our arguments are generalizable to other concepts, such as that of free will.

In constructing or criticizing theories, philosophers have appealed to intuitive judgments in thought experiments, as seen in discussions of epistemology (Gettier, 1963), philosophy of mind (Searle, 1980; Chalmers, 1996), philosophy of language (Kripke, 1980), and so on. In doing so, philosophers have appealed to the fact that their intuitive judgments in thought experiments reflect their own theories. However, experimental philosophy has found that people's intuitive judgments are affected by unexpected factors, beyond those conceived by philosophers.

One famous example is intuitive judgments on the intentionality of action (Knobe, 2003). Participants were presented with a story about a person working in a company. The story indicates that the person is working on a new business program and knows that its side effect will harm the environment. However, the person says that he does not care about this outcome and carries out the program, which leads to the bad side effect ("Harm" condition). Another version of the story is the same, except that the side effect of the program is good for the environment ("Help" condition). The participants read one of these stories and were asked whether the person has intentionally harmed (helped) the environment. Surprisingly, a strong asymmetry was found between the two conditions: most participants in the Harm condition responded that his action was intentional, while most in the Help condition replied that it was not intentional, despite the identical structure of the stories. Thus, people's intuitive judgments on the intentionality of an action vary, according to the perceived harmfulness/helpfulness of the action. It seems that, generally speaking, we attribute intentionality to harmful or morally bad side effects. This tendency of attribution is called the "Knobe effect."

The Knobe effect may reflect the concept of intentionality. Intentionality attribution, which is the crucial function of our "theory of mind" (ToM), is partly driven by moral cognition in nature and is thus the result of the appropriate application of the concept of intentionality. In this case, the concept of intentionality can be regarded as being constituted by moral cognition. However, the Knobe effect may reflect the inappropriate interference of moral cognition that is not constitutive of the concept of intentionality (Nadelhoffer, 2004, 2006). This question about the nature of the Knobe effect has often been framed in terms of *competence* and *performance*. The Knobe effect reflects the competence of the concept of intentionality (i.e., the core of ToM) or an error in the performance of it (Alexander, 2012).

The distinction between competence and performance originates in generative linguistics. Chomsky (1965, p. 3) argues that competence is linguistic knowledge that is possessed by an ideal

speaker-listener in a language community. Because the actual linguistic performance is affected by a variety of constraints, such as memory capacities, attention controls, vocal functions, and so on, language competence is perfectly reflected in performance only in the idealization of these functions. This distinction has yielded noteworthy results by factoring out heterogeneous, confounding factors from the main target of linguistic theories (see discussion by Jackendoff, 2002).

There is an ongoing debate on how to characterize psychologically the relationships between competence and performance (Phillips, 2004; Marantz, 2005; Neeleman and van de Koot, 2010; Phillips and Lewis, 2013). Neeleman and van de Koot (2010) argue that competence and performance should be understood as theories of the same language system but at different descriptive levels. In this case, competence and performance would roughly correspond to different levels of analysis, i.e., the computational and algorithmic levels introduced by Marr (1982). However, we will not discuss this issue further in this paper, and we hereafter use the term competence/performance in Chomsky's (1965) original sense, which seems to be dominant in the literature in experimental philosophy.

The distinction between competence and performance can be applied to concepts. The Knobe effect may reflect the competence of the concept of intentionality, such that moral cognition underlies the application of the concept. For example, Knobe (2006, p. 226) states that "moral considerations are playing a helpful role in people's underlying competence itself." Otherwise, the Knobe effect is a sort of performance error, where moral cognition distorts the application of the concept. Generally speaking, when a judgment involving a concept is affected by some psychological factor, it may reflect the competence of the concept, or it may be the result of error in its performance.

We can also pose this type of question in relation to many other studies. For instance, people's intuitive judgments about free will have been examined, focusing on whether the concept of free will is compatible with determinism. Some scholars argue that the concept of free will is compatibilist, since the participants attribute free will to fictional characters in a deterministic world (Nahmias et al., 2005). However, other studies suggest that people's judgments are sensitive to whether they are presented with abstract or concrete scenarios. People attribute free will in concrete scenarios much more than they do in abstract ones (Nichols and Knobe, 2007; cf. De Brigard et al., 2009; Mandelbaum and Ripley, 2012). Although this may be the case, an issue regarding the nature of the concept of free will remains to be resolved. Even if our attribution of free will is affected by the perception of concreteness, it is unclear whether such an effect reflects the competence of the concept of free will.

Interestingly, some scholars seem to suggest that the nature of people's concepts has philosophical implications. When discussing how to interpret the Knobe effect, Adams and Steadman (2004, p. 173) mention the philosophical view that intentionality does not require intention, which the Knobe effect "may be taken to support." In discussing free will, Nahmias et al. (2006, p. 30) maintain that "[b]ecause the free will debate is intimately connected to ordinary intuitions and beliefs via these values and practices, it is important that a philosophical theory of free will accounts for and accords with ordinary people's understanding of the concept and their judgments about relevant cases." In discussing the general background of experimental philosophy, Knobe and Nichols (2008b, p. 12) state, "[m]ore and more, philosophers are coming to feel that questions about how people ordinarily think have great philosophical significance in their own right." Indeed, these scholars often seek to grasp the concepts of intentionality and free will in terms of competence/performance.

Here two questions arise. First, do experimental results, such as the Knobe effect, reflect the competence of the intentionality concept or a mere performance error? Second, how and why does such an understanding inform philosophical debates about intentionality? In what follows, we consider these two questions in turn.

DEVELOPMENTAL AND DISABILITY STUDIES

Practices in linguistics may provide a clue in distinguishing between competence and performance. In linguistics, when competing theories possess identical explanatory powers, the possibility of language acquisition has been successfully used to constrain the range of theory (Chomsky, 1965; Yang, 2010). Moreover, agrammatism, which is a type of aphasia specific to syntactic processing, has been useful in clarifying domain-specific linguistic competence by dissociating domain-general components (Friedmann and Grodzinsky, 1997; Kinno et al., 2009). In a similar way, developmental and disability studies of the concept of intentionality may also help us to theorize the nature of the concept.

First, the cognitive nature of the concept of intentionality may be clarified by considering developmental studies of the Knobe effect. According to the experimental study of Leslie et al. (2006a), children as young as four showed the same tendency as adults. Moreover, the Knobe effect appeared as soon as the children learned to understand the concept of "do not care [bad side effects]" that was included in the experiment's scenario. These results suggest that our innate concept of intentionality grows and fits with the Knobe effect. Segal (2008) argues, "it is difficult to believe that they learned it from observation of adult patterns of judgment, or that they inferred it from something else. It looks as though this is just how FP [folk psychology] grows" (ibid, p. 101). Thus, the Knobe effect is essentially associated with the concept of intentionality within ToM and reflects the competence of this concept.

Second, the cognitive nature of the concept of intentionality may be further clarified by considering people with autism spectrum disorder, who generally show some impairment in the ToM. On the autism spectrum, adults with Asperger's syndrome or high-functioning autism (hereafter, AS/HFA) have no apparent disabilities in general intelligence and language and can generally pass a simple false-belief task, which is a simple test of the ToM. Moreover, with regard to basic moral perception, several studies have shown that there is no large difference between people on the autism spectrum and people with typical development (Blair, 1996; Grant et al., 2005; Leslie et al., 2006b). People with AS/HFA have difficulties, however, in understanding the mental states of others in complex situations and in passing higher level ToM tests, which involve sarcasm, irony, bravado, and the like. From these observations, it has been suggested that adults with

AS/HFA use heuristics, which differ from the core of ToM, in order to understand the minds of others (Happé, 1995; Tager-Flusberg and Joseph, 2003). Thus, people with AS/HFA have impairments in the core of the ToM and use specific heuristics for the attribution of intentionality to complement these impairments, while their basic moral perceptions are normal.

A hypothesis regarding the Knobe effect in people with AS/HFA can be derived from these assumptions. If the Knobe effect reflects the competence of the concept of intentionality (i.e., the core of the ToM), we will not observe it in people with AS/HFA who have impairments in the core. Otherwise, if the Knobe effect is not a manifestation of the core of the ToM but is at best only a sign of heuristics in intentionality attribution, we will observe it in people with AS/HFA. In the light of the current empirical literature, the latter possibility is likely. A recent study of the Knobe effect in the autism spectrum group has revealed that even people with AS/HFA show the Knobe effect similarly to people with typical development (Zalla and Leboyer, 2011). Although further studies were needed, in this case, the Knobe effect would be regarded as a sign of heuristics in intentionality attribution and therefore irrelevant to the nature of the concept of intentionality.

As shown above, disability and developmental studies offer us a theoretical advance in understanding the nature of the Knobe effect. At the moment, there is conflicting evidence from such studies regarding whether the Knobe effect reflects competence.

NORMATIVIST PSYCHOLOGY AND EXPERIMENTAL PHILOSOPHY

Here, we question the implications of the above findings. What are the philosophical implications of the fact that the Knobe effect reflects competence or performance error? One related philosophical problem is the relationship between intentionality and *intention*. While it seems that intention is required for an action to be intentional, some philosophers reject this conclusion (cf. Bratman, 1984; Adams, 1986). If the Knobe effect exactly reflects the competence of the intentionality concept, then the idea of disconnecting intentionality and intention may be supported, since intentionality could be attributed to the side effects of an action that were not intended (and such an attribution reflects the competence of the concept of intentionality). However, we would like to point out that an assumption is required for this philosophical implication.

The assumption is that the distinction between competence and performance error implies the distinction between the rationality and irrationality of concepts. Generally speaking, when a particular intuitive judgment cannot be regarded as rational, a philosophical argument based on this judgment is not justified, even if the intuitive judgment reflects the competence of the related concept. For example, when the Knobe effect does not reflect a rational thought, philosophical arguments on intentionality need not necessarily take the Knobe effect into account, even if the concept of intentionality is constituted by moral cognition. Here, we tentatively characterize the rationality of thought, including judgments and reasoning, as the disposition to produce true beliefs. This type of rationality, which has been regarded as essential in philosophical discussions, is called *epistemic* rationality, and it is distinct from other types of rationality, such as instrumental or ecological rationality. As long as our main concern is how we *should* think about the nature of intentionality, than we *do* think about it; there is no reason for philosophical theories to take into account whether the Knobe effect reflects competence, since it does not guarantee epistemic rationality.

Here, we can follow Elqayam and Evans (2011) in making a distinction between *descriptivism* and *normativism* in psychology. In general, normativist psychology directly relates competence and performance error to rational and irrational thought, respectively. By assuming particular norms, normativist psychology classifies thoughts as rational or irrational, depending on whether the nature of the thoughts accords with certain norms, and judges the distinction between the competence and performance error of the thoughts. Elqayam and Evans (2011) argue that psychology should follow linguistics, which adopts descriptivism and proposes a theory regarding the competence of language. In other words, psychology should dedicate itself to describe competence in accordance with descriptivism. They claim that psychology may distinguish competence and performance, but it should not engage in judging whether thoughts are rational or irrational. Otherwise, it draws *ought* from *is*, whose inference is generally unsupported.

In the same way, we want to point out that the distinction between competence and performance error has philosophical implications, only when we adopt normativism and assume the following norm: To be rational in the sense of being disposed to produce true beliefs, thoughts should reflect the competence of concepts. Indeed, if we assume this norm, we obtain the following consequence: If the Knobe effect reflects competence, it can be regarded as a *rational judgment* and thus should be considered in philosophical arguments. However, we suggest that how we *ought to* think about the nature of intentionality has to be distinguished from how typically developed people *do* think about it in everyday life.

Let us consider the above research on disabilities of the ToM where people with AS/HFA evince the Knobe effect. From this fact, we might infer that the Knobe effect reflects the heuristics but not the competence of the concept of intentionality (the core of the ToM). However, this view alone does not lead to the conclusion that the Knobe effect is not a rational judgment. In order to obtain such a conclusion, we must also assume that judgments in accordance with the functioning of the core can be regarded as rational. This assumption is a normative assertion regarding human thoughts. On what ground do thoughts based on the core have priority in terms of rationality? The fact that they were acquired in typical development cannot offer the reason, since the reason why typical development is directed to rational thought is questioned here. Whether the Knobe effect reflects competence is not directly relevant to truths about intentionality, since one cannot infer norms, at least in psychological facts, that make some judgments rational and others irrational.

Thus, we cannot draw philosophical implications without normativism, even while we can identify the cognitive concepts in terms of competence and performance. This requirement of normativism has not been explicated in the debates within experimental philosophy. While the importance of whether the Knobe effect reflects the competence of the concept of intentionality has often been suggested, the idea of competence/performance does

not guarantee anything like (ir)rationality. Notice that this is also the case in linguistic studies, where the idea of competence is a kind of biological function, which implies nothing about epistemic rationality. This is true regardless of the general characterization of biological functions. Even if our moral cognition has some causal role in the application of the concept of intentionality or if the interaction between moral cognition and intentionality attribution is important for biological adaptation, it has nothing to do with how often we arrive at the truth about intentionality (cf. Stich, 1990). Therefore, we conclude that there is a large gap between the idea of competence, which is a kind of biological function, and (ir)rationality.

CONCLUSIONS

We suggest that it is empirically possible to determine whether the application of the concept of intentionality reflects its competence or error in performance. However, we point out that this fact about competence/performance does not imply anything positive about the nature of intentionality, contrary to the assumption made by some scholars in experimental philosophy. Drawing these implications from competence/performance requires a normativist psychology, which we think is a doubtful methodology. Thus, we have to bring something from outside psychology to bridge the gap between the competence/performance and (ir)rationality of concepts. For example, we might suggest that the concept of intentionality works for social interaction (cf. Knobe, 2006) and thus that socially admitted norms about its use should be reflected in any theory of intentionality. Otherwise, we might suggest a constitutive approach in general, which claims that our intuitions produce conceptual truth "by drawing on constructs such as reflective equilibrium and constitutive norms" (Evans and Elqayam, 2011, p. 283; cf. Thomasson, 2012). While we do not believe that experimental results do not inform philosophy, it seems better to explore something beyond the competence/performance of concepts. In any case, we have to be aware of the gap between how we should think and how we think about intentionality.

ACKNOWLEDGMENTS

This paper is largely based on our previous work, Iijima and Ota (2014). We appreciate *Frontiers in Psychology* for publishing this paper and *The Nagoya Journal of Philosophy* for admitting it. This work was supported by a Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Fellows (26-11465) and the Suntory Foundation.

REFERENCES

Adams, F. (1986). Intention and intentional action: the simple view. *Mind Lang.* 1, 281–301. doi: 10.1111/j.1468-0017.1986.tb00327.x

Adams, F., and Steadman, A. (2004). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis* 64, 173–181. doi: 10.1111/j.1467-8284.2004.00480.x

Alexander, J. (2012). *Experimental Philosophy: An Introduction.* Malden: Polity.

Blair, R. J. R. (1996). Brief report: morality in the autistic child. *J. Autism Dev. Disord.* 26, 571–579. doi: 10.1007/BF02172277

Bratman, M. (1984). Two faces of intention. *Philos. Rev.* 93, 375–405. doi: 10.2307/2184542

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory.* Oxford: Oxford University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge: The MIT Press.

De Brigard, F., Mandelbaum, E., and Ripley, D. (2009). Responsibility and the brain sciences. *Ethical Theory Moral Pract.* 12, 511–524. doi: 10.1007/s10677-008-9143-5

Elqayam, S., and Evans, J. St. B. T. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Evans, J. St. B. T., and Elqayam, S. (2011). Towards a descripitivist psychology of reasoning and decision making. *Behav. Brain Sci.* 34, 275–290. doi: 10.1017/S0140525X11001440

Friedmann, N., and Grodzinsky, Y. (1997). Tense and agreement in agrammatic production: pruning the syntactic tree. *Brain Lang.* 56, 397–425. doi: 10.1006/brln.1997.1795

Gettier, E. (1963). Is justified true belief knowledge? *Analysis* 23, 121–123. doi: 10.1093/analys/23.6.121

Grant, C. M., Boucher, J., Riggs, K. J., and Grayson, A. (2005). Moral understanding in children with autism. *Autism* 9, 317–331. doi: 10.1177/1362361305055418

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Dev.* 66, 843–855. doi: 10.2307/1131954

Iijima, K., and Ota, K. (2014). Competence and rationality: what does experimental philosophy have beyond psychological implications? [Japanese]. *Nagoya J. Philos.* 11, 39–61.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198270126.001.0001

Kinno, R., Muragaki, Y., Hori, T., Maruyama, T., Kawamura, M., and Sakai, K. L. (2009). Agrammatic comprehension caused by a glioma in the left frontal cortex. *Brain Lang.* 110, 71–80. doi: 10.1016/j.bandl.2009.05.001

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis* 63, 190–194. doi: 10.1111/1467-8284.00419

Knobe, J. (2006). The concept of intentional action: a case study in the uses of folk psychology. *Philos. Stud.* 130, 203–231. doi: 10.1007/s11098-004-4510-0.

Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., and Sommers, T. (2012). Experimental philosophy. *Annu. Rev. Psychol.* 63, 81–99. doi: 10.1146/annurev-psych-120710-100350

Knobe, J., and Nichols, S. (eds). (2008a). *Experimental Philosophy.* Oxford: Oxford University Press.

Knobe, J., and Nichols, S. (2008b). "An experimental philosophy manifesto," in *Experimental Philosophy*, eds J. Knobe and S. Nichols (Oxford: Oxford University Press), 3–14.

Knobe, J., and Nichols, S. (eds). (2013). *Experimental Philosophy*: Vol. 2. Oxford: Oxford University Press.

Kripke, S. A. (1980). *Naming and Necessity.* Cambridge: Harvard University Press.

Leslie, A. M., Knobe, J., and Cohen, A. (2006a). Acting intentionally and the side-effect effect: theory of mind and moral judgment. *Psychol. Sci.* 17, 421–427. doi: 10.1111/j.1467-9280.2006.01722.x

Leslie, A. M., Mallon, R., and Dicorcia, J. A. (2006b). Transgressors, victims, and cry babies: is basic moral judgment spared in autism? *Soc. Neurosci.* 1, 270–283. doi: 10.1080/17470910600992197

Mandelbaum, E., and Ripley, D. (2012). Explaining the abstract/concrete paradoxes in moral psychology: the NBAR hypothesis. *Rev. Philos. Psychol.* 3, 351–368. doi: 10.1007/s13164-012-0106-3

Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *Linguist. Rev.* 22, 429–445. doi: 10.1515/tlir.2005.22.2-4.429

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W. H. Freeman.

Nadelhoffer, T. (2004). On praise, side effects, and folk ascriptions of intentionality. *J. Theor. Philos. Psychol.* 24, 196–213. doi: 10.1037/h0091241

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: some problems for juror impartiality. *Philos. Explor.* 9, 203–219. doi: 10.1080/13869790600641905

Nahmias, E., Morris, S. G., Nadelhoffer, T., and Turner, J. (2005). Surveying freedom: folk intuitions about free will and moral responsibility. *Philos. Psychol.* 18, 561–584. doi: 10.1080/09515080500264180

Nahmias, E., Morris, S. G., Nadelhoffer, T., and Turner, J. (2006). Is incompatibilism intuitive? *Philos. Phenomenol. Res.* 73, 28–53. doi: 10.1111/j.1933-1592.2006.tb00603.x

Neeleman, A., and van de Koot, H. (2010). "Theoretical validity and psychological reality of the grammatical code," in *The Linguistics Enterprise: From Knowledge of Language to Knowledge in Linguistics*, eds M. Everaert, T. Lentz, H. de Mulder, Ø. Nilsen, and A. Zondervan (Amsterdam: John Benjamins), 183–212.

Nichols, S., and Knobe, J. (2007). Moral responsibility and determinism: the cognitive science of folk intuitions. *Noûs* 41, 663–685. doi: 10.1111/j.1468-0068.2007.00666.x

Phillips, C. (2004). "Linguistics and linking problems," in *Developmental Language Disorders: From Phenotypes to Etiologies*, eds M. Rice and S. Warren (Mahwah: Lawrence Erlbaum Associates), 241–287.

Phillips, C., and Lewis, S. (2013). Derivational order in syntax: evidence and architectural consequences. *Stud. Linguist.* 6, 11–47.

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424. doi: 10.1017/S0140525X00005756

Segal, G. (2008). "Poverty of stimulus arguments concerning language and folk psychology," in *The Innate Mind, Volume 3, Foundations and the Future*, eds P. Carruthers, S. Laurence, and S. Stich (Oxford: Oxford University Press), 90–105.

Stich, S. (1990). *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation.* Cambridge: The MIT Press.

Tager-Flusberg, H., and Joseph, R. M. (2003). Identifying neurocognitive phenotypes in autism. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 303–314. doi: 10.1098/rstb.2002.1198

Thomasson, A. L. (2012). Experimental philosophy and the methods of ontology. *Monist* 95, 175–199. doi: 10.5840/monist201295211

Yang, C. D. (2010). Three factors in language variation. *Lingua* 120, 1160–1177. doi: 10.1016/j.lingua.2008.09.015

Zalla, T., and Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. *Rev. Philos. Psychol.* 2, 681–698. doi: 10.1007/s13164-011-0048-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 February 2014; accepted: 07 July 2014; published online: 22 July 2014.
Citation: Iijima K and Ota K (2014) How (not) to draw philosophical implications from the cognitive nature of concepts: the case of intentionality. Front. Psychol. ***5****:799. doi: 10.3389/fpsyg.2014.00799*
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 16 June 2014
doi: 10.3389/fpsyg.2014.00598

Against a normative view of folk psychology

Meredith R. Wilkinson*

Division of Psychology, Faculty of Health and Life Sciences, School of Applied Social Sciences, De Montfort University, Leicester, UK
**Correspondence: mwilkinson@dmu.ac.uk*

Edited by:
David E. Over, Durham University, UK

Reviewed by:
Linden John Ball, University of Central Lancashire, UK

Keywords: folk psychology, normative, reasoning, descriptive, theory theory, simulation theory

Recently Elqayam and Evans (2011) have proposed that researchers studying human thinking should be moving away from normative accounts that specify how we "ought" to reason to a more descriptivist framework that describes *how* we reason. This is an approach that I very much support. The aim of the present article is to demonstrate how this can be applied to the study of mental state reasoning in terms of folk psychology (FP). Folk psychology refers to our everyday ability to attribute mental states to other people, including their beliefs, desires, intentions and so forth (e.g., Ratcliffe and Hutto, 2007). I do not want to deny that FP *can* be normative. Indeed, there are many instances where normative responding is required. For example, in the traditional false belief task (e.g., Baron-Cohen et al., 1985) there is a "right" or "wrong" answer - a single norm paradigm (Elqayam and Evans, 2011). However, it may be the case that FP is normative in certain circumstances (e.g., the false belief task) but as I shall suggest below this is not always the case. By viewing FP as normative what researchers end up doing is ignoring the *processes* of how such inferences arise. What I want to propose is that viewing FP as normative is problematic since it reduces mental state inferences to simply being "right" or "wrong." I propose that moving away from a normative agenda in FP and embracing a more descriptivist framework proves extremely useful for our understanding of how we understand others' minds.

FOLK PSYCHOLOGY AS A NORMATIVE FRAMEWORK

One factor that philosophers have proposed regarding FP is that it is normative in nature. One of the earliest claims of this was made by Dennett (1989) who argued:

"Folk psychology, then, is idealized in that it produces its predictions and explanations by calculating in a normative system; it predicts what we will believe, desire, and do, by determining what we ought to believe, desire, and do" (Dennett, 1989 p. 52).

Dennett (1989) views FP as a form of mindreading but has the perspective of the intentional stance. His perspective appears to be normative in nature specifying both a normative framework and an ought stance. This sense of normativism in FP extends to recent literature:

"Even on the standard view, then, folk psychology is not just an explanatory/predictive practice, it is also, in a sense, a normative practice: a practice of showing how people's performances lives up to certain norms and thereby become, in that special way, intelligible. Although folk psychologists may have some context-specific views about what others will do—based, for instance, on experience—the bulk of these views will be heavily influenced by norm-governed judgments about what others ought to do, what it makes sense to do in the circumstances" (McGeer, 2007, p. 141).

The above quotation appears to be arguing that we can in our FP responses generate normative responses. Just as we have a normative rule that when driving you should stop at a red light, what is being implied here is that we have a sense of what people "ought" to do in certain situations. Viewing FP in a normative framework is a view that exists till the present day:

"Whatever focus one adopts, judgments of the rational or scientific status of elements in folk psychology are inevitably normative judgments, based on comparisons between what ordinary folk do with some prescriptive scientific account" (Fletcher, 1995, pp. 43–44).

Such a perspective sees FP as confirming to rationality and having an analogy with science. However, if FP is to have an analogy with science then we may to some degree want to subject it to empirical testing. However, Churchland (1991) argues that empirical testing may do little for FP:

"Folk psychology, insist others is radically unlike the examples cited. It does not consist of laws. It does not support causal explanations. It does not evolve over time. Its central purpose is normative rather than descriptive. And thus, it is not the sort of framework that might be shown to be radically defective by sheerly empirical findings" (Churchland, 1991, p. 51).

I think that this quote is somewhat problematic as FP does support causal explanations and it has evolved over time. For example, the use of neuroscience to examine FP (e.g., Ruby and Decety, 2001). I hope to have demonstrated in this section how multiple theorists view FP as normative and now aim to demonstrate what is problematic about doing so.

WHAT IS PROBLEMATIC ABOUT VIEWING FOLK PSYCHOLOGY AS NORMATIVE?

Whilst I accept that there is normative responding in FP, for example, it is normative to assume that if we push someone off a seat on the bus so that we can sit down then they will be angry I believe that viewing FP as normative is problematic since it reduces mental state reasoning to "right" or "wrong" answers and indeed a "right" way to reason (as indicated by the quotations above). Admittedly, FP is

a single norm paradigm (Elqayam and Evans, 2011) so does not face the difficulty that reasoning research does of having multiple normative accounts to arbitrate between.

I agree with Elqayam and Evans (2011) when they claim that normativism has biased the study of thinking and feel that this can be applied to the study of FP. FP has repeatedly made use of tasks of false belief in order to examine mentalizing. I believe that this is problematic since it has led to a very restricted range of tasks being studied. If we were to move away from a normative perspective of FP then this would open many more doors to examine mentalizing since far too much attention, from my perspective, has been focused on the false belief paradigm. Thus, what is happening here is that people are either attributed with having a capacity to engage in FP reasoning or not. I believe that this is problematic since there is much more to FP then the false belief task and much more to the false belief task than FP understanding (Bloom and German, 2000). I demonstrate within the next section how a descriptivist study of FP may work.

A final problem with viewing FP as normative is that although there are clear cut cases, as in the false belief task, where there is a "right" and "wrong" answer in tasks of mental state reasoning this isn't always going to be the case. I believe that if something is to be *fully* normative then it should always be the case that there is a right and wrong response. For example, if we are informed by our friend that her boyfriend has ended their relationship we may assume that she will be devastated. However, other factors may influence that judgment such as if she wanted to separate with him then you may believe that she will be relieved. However, it is still possible that she will be upset as he separated with her first. What I aim to demonstrate here is that there is not always a clear cut answer with FP and therefore we should embrace a more descriptive framework rather than a normative one.

WHAT ARE THE ADVANTAGES OF A DESCRIPTIVIST ACCOUNT?

I have argued that viewing FP as normative can be severely problematic. I am not the only theorist who takes this view. Andrews (2012) argues that "the study of folk psychology is a descriptive endeavor, as opposed to a normative one." (Andrews, 2012, p. 251). Thus, what we should be aiming to do as researchers is not provide an account of how people *ought* to reason but provide an account of how they do reason using appropriate theories and experimental methodologies.

A recent descriptivist approach to FP comes from Wilkinson and Ball (2013) who provide a dual-process perspective to theorizing and simulation. I draw the link to the theorizing vs. simulation theory debate here since it is the ideal type of question that those who study FP should be asking in terms of the cognitive processes which underlie FP, I am not saying that the term FP necessarily refers to the theorizing vs. simulation debate itself. Theorizing refers to understanding mental states via the adoption of theories which link mental states, behavior and the environment together (e.g., Carruthers, 1996). Whereas simulation proposes that we reason about others' mental states via controlled processes of simulation either from a third-person (e.g., Goldman, 2006) or first-person (e.g., Gordon, 1986) perspective. According to Wilkinson and Ball theorizing is viewed as synonymous with intuitive reasoning and simulation is viewed as synonymous with reflective reasoning within a dual-process framework (e.g., Evans, 2010). As such, theorizing can be viewed as possessing the characteristics of being fast, automatic, low effort, high capacity and independent of working memory resources whereas simulation is slow, controlled, high effort, low capacity and dependent upon working memory resources.

According to Wilkinson and Ball (2013) people can either choose to theorize or simulate. It is possible for them to engage in both with people skipping between theorizing and simulation and vice versa. This reflects the hybrid nature of theorizing and simulation (see also Mitchell et al., 2009). Wilkinson and Ball note that conflict may arise between the responses generated by theorizing and simulation and this can be overcome with a conflict resolution mechanism which is analogous to Evans' (2009) type 3 reasoning. Conflict does not have to arise though and one response may just be generated. The advantage of viewing FP in this manner is that it promotes a program of research which focuses on the question of 'how' people reason and not just how they ought to reason. It further avoids the tricky issue of rationality, something which in mental state reasoning would be very difficult to examine since what is rational for one agent is not necessarily rational for another. I believe that this addresses the issue raised above regarding how viewing FP as normative has led to a bias in how it is studied. Wilkinson et al. (2010) used think aloud protocols where they asked participants to think aloud whilst working through regret-orientated counterfactual scenarios and then coding participants' verbalizations for instances of theorizing and simulation. Adopting this method enabled Wilkinson et al. (2010). To gain a measure of *how* people reason which gives a much richer insight than whether someone answered correctly or incorrectly.

Wilkinson and Ball (2013) are not the only researchers to link FP to dual-process theories. Bohl and van den Bos (2012) propose that theory of mind requires reflective reasoning whereas interactionism requires intuitive reasoning. Whilst their account differs from Wilkinson and Ball since they propose that theory of mind consists of both intuitive and reflective processes both demonstrate a move toward viewing mental state reasoning in dual-process terms (see also Apperly and Butterfill, 2009).

I, like Andrews (2012), believe that FP is descriptive rather than normative in nature and Andrews aims to develop an account which is more descriptively accurate than normative. It is only via viewing FP as descriptive can real progress be made into examining the complexities of our abilities to engage in mental state reasoning of both ourselves and other people. Within the reasoning literature some authors endorse a "soft normativsm" perspective (e.g., Stupple and Ball, 2011) whereby they propose that normative constructs can feed into a descriptivist framework. To some degree I endorse this perspective since there are normative rules which govern how we expect others to feel. However, I do believe that the descriptivist framework of Wilkinson and Ball (2013) enables a much richer account of the cognitive processes in mental state reasoning than any normative only perspective permits.

ACKNOWLEDGMENTS

I would like to thank Linden Ball and Rachel Cooper for many interesting discussions during my PhD thesis on folk psychology. I would also like to thank Shira Elqayam for many valued discussions on reasoning more generally together with her support and encouragement in writing this paper.

REFERENCES

Andrews, K. (2012). *Do Apes Read Minds? Toward a New Folk Psychology*. Cambridge, MA: MIT Press.

Apperly, I. A., and Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychol. Rev.* 116, 953–970. doi: 10.1037/a0016923

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8

Bloom, P., and German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77, B25–B31. doi: 10.1016/S0010-0277(00)00096-2

Bohl, V., and van den Bos, W. (2012). Toward an integrative account of social cognition: marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Front. Hum. Neurosci.* 6:274. doi: 10.3389/fnhum.2012.00274

Carruthers, P. (1996). "Autism as mind-blindness: an elaboration and partial defence," in *Theories of Theories of Mind*, eds P. Carruthers and P. K. Smith (Cambridge: Cambridge University Press), 257–273.

Churchland, P. M. (1991). "Folk psychology and the explanation of human behavior," in *The Future of Folk Psychology: Intentionality and Cognitive Science*, ed J. D. Greenwood (Cambridge, UK: Cambridge University Press), 51–69.

Dennett, D. C. (1989). *The Intentional Stance*. Cambridge, MA: MIT press.

Elqayam, S., and Evans, J. S. B. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Evans, J. St. B. T. (2009). "How many dual-process theories do we need? One, two, or many?," in *Two minds: Dual Processes and Beyond*, eds J. St. B. T. Evans and K. Frankish (Oxford: Oxford University Press), 33–54.

Evans, J. St. B. T. (2010). *Thinking Twice: Two Minds in One Brain*. Oxford: Oxford University Press.

Fletcher, G. J. (1995). *The Scientific Credibility of Folk Psychology Lawrence Erlbaum*. Hillsdale, NJ: Psychology press.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York, NY: Oxford University Press. doi: 10.1093/0195138929.001.0001

Gordon, R. M. (1986). Folk psychology as simulation. *Mind Lang.* 1, 158–171. doi: 10.1111/j.1468-0017.1986.tb00324.x

McGeer, V. (2007). "The regulative dimension of folk psychology," in *Folk Psychology Re-Assessed*, eds D. Hutto and M. Ratcliffe (Dordrecht: Springer), 137–156. doi: 10.1007/978-1-4020-5558-4_8

Mitchell, P., Currie, G., and Ziegler, F. (2009). Two routes to perspective: simulation and rule−use as approaches to mentalizing. *Br. J. Dev. Psychol.* 27, 513–543. doi: 10.1348/026151008X334737

Ratcliffe, M., and Hutto, D. D. (2007). "Introduction" in *Folk Psychology Re-Assessed*, eds D. D. Hutto and M. Ratcliffe (Dordrecht: Springer), 1–22. doi: 10.1007/978-1-4020-5558-4_1

Ruby, P., and Decety, J. (2001). Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nat. Neurosci.* 4, 546–550. doi: 10.1038/87510

Stupple, E. J., and Ball, L. J. (2011). Normative benchmarks are useful for studying individual differences in reasoning. *Behav. Brain Sci.* 34, 270–271. doi: 10.1017/S0140525X11000562

Wilkinson, M. R., and Ball, L. J. (2013). "Dual processes in mental state understanding: is theorising synonymous with intuitive thinking and is simulation synonymous with reflective thinking?" in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, eds M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Austin, TX: Cognitive Science Society), 3771–3776.

Wilkinson, M. R., Ball, L. J., and Cooper, R. (2010). Arbitrating between Theory-Theory and Simulation Theory: evidence from a think-aloud study of counterfactual reasoning," in *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society* (Austin, TX), 1008–1013.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 February 2014; accepted: 28 May 2014; published online: 16 June 2014.

Citation: Wilkinson MR (2014) Against a normative view of folk psychology. Front. Psychol. ***5****:598. doi: 10.3389/fpsyg.2014.00598*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Evolutionary and ecological accounts

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 15 May 2014
doi: 10.3389/fpsyg.2014.00435

The nature of thinking, shallow and deep

Gary L. Brase*

Department of Psychological Sciences, Kansas State University, Manhattan, KS, USA

Edited by:
David E. Over, Durham University, UK

Reviewed by:
Ingmar Visser, Universiteit van Amsterdam, Netherlands
Meredith Ria Wilkinson, De Montfort University, UK

****Correspondence:***
Gary L. Brase, Department of Psychological Sciences, Kansas State University, 492 Bluemont Hall, Manhattan, KS 66506, USA
e-mail: gbrase@ksu.edu

Because the criteria for success differ across various domains of life, no single normative standard will ever work for all types of thinking. One method for dealing with this apparent dilemma is to propose that the mind is made up of a large number of specialized modules. This review describes how this multi-modular framework for the mind overcomes several critical conceptual and theoretical challenges to our understanding of human thinking, and hopefully clarifies what are (and are not) some of the implications based on this framework. In particular, an evolutionarily informed "deep rationality" conception of human thinking can guide psychological research out of clusters of *ad hoc* models which currently occupy some fields. First, the idea of deep rationality helps theoretical frameworks in terms of orienting themselves with regard to time scale references, which can alter the nature of rationality assessments. Second, the functional domains of deep rationality can be hypothesized (non-exhaustively) to include the areas of self-protection, status, affiliation, mate acquisition, mate retention, kin care, and disease avoidance. Thus, although there is no single normative standard of rationality across all of human cognition, there are sensible and objective standards by which we can evaluate multiple, fundamental, domain-specific motives underlying human cognition and behavior. This review concludes with two examples to illustrate the implications of this framework. The first example, decisions about having a child, illustrates how competing models can be understood by realizing that different fundamental motives guiding people's thinking can sometimes be in conflict. The second example is that of personifications within modern financial markets (e.g., in the form of corporations), which are entities specifically constructed to have just one fundamental motive. This single focus is the source of both the strengths and flaws in how such entities behave.

Keywords: normative models, cognitive modularity, deep rationality, evolutionary psychology, human reasoning, time scales in rational decision making

INTRODUCTION

> *It has been said that man is a rational animal. All my life I have been searching for evidence which could support this.*
>
> Russell (1950)

A foundational principle of science is that good scientific theories must generate testable, and therefore falsifiable, hypotheses. Research can then obtain data relevant to that hypothesis and show that the prediction either holds up or fails. Science thus has normative standards as an intrinsic property of the scientific method; the nature of good scientific theories includes the ability to provide testable predictions. Those predictions provide a standard for evaluating a theory, saying what ought to happen if the theory is correct. In this sense, then, normative models are an essential property of research on human though (or research on anything else). Certainly people can become confused between theoretical predictions of what "ought" to happen (in the sense of as the theory predicts) versus notions of "ought" based on a cultural or socio-moral position. That, however, is not a problem with normative standards as a property of scientific theories as much as it is a problem of people not understanding how science works. So, for example, when one asks "what *ought* human thinking be like?" it is important to clarify if the question is about a prediction based on a scientific theory or if the question carries some presumption of the inquisitor based on their personal views. One of these is science; the other is personal opinion.

The problem with normative models in the scientific study of human thought is that no single normative standard works for all types of thinking. How do we decide on appropriate normative standards? (Which, in this scientific sense means how do we decide upon appropriate theoretical frameworks?) Thinking is a ubiquitous feature of human activity, but the normative standards for evaluating good food are different from the normative standards for evaluating a good place to live, which are different from the normative standards for evaluating a good relationship partner, which are in turn different from the normative standards for evaluating a good stock market decision. In general terms, for any problem or task domain there is a set of features that define that problem/task and therefore those same features provide criteria for success (i.e., the "good" solution). The more one knows about the nature of the features that constitute a problem, the more one therefore knows about properties that can be exploited to get to an effective and efficient solution. For instance, some of the defining features of the problem of food acquisition are identifying high calorie, digestible items. The criteria for success ("good" food) are things which contain fats, sugars, carbohydrates, and

proteins. Other things (dirt, wood, metal, plastic) do not satisfy these criteria. If one attempts to collapse across multiple problems or tasks to achieve a general-purpose solution method, the features that define the overall problem become increasingly general and computationally ineffectual. At moderate levels of generality we find problem solving tools that are simply weak (e.g., General Problem Solver; Newell and Simon, 1972). With further levels of generality we find only normative standard that are uselessly vague ("Don't screw up.") and computational incapacitation as a result.

Because different normative models provide standards of evaluation for different types, or domains, of behaviors, one of the key questions then is how to parse the various aspects of the world into domains. In which domains do which particular normative models apply? Some people will recognize this as the dilemma posed by the idea (from Plato's *Phaedrus*) that scientific theories should "carve nature at its joints," but the problem is that there does not appear to be any single carving pattern that consistently and uniquely works. Instead there seem to be multiple carving patterns that can each be legitimately argued for and that each nonetheless have flaws. In other words, even within a particular domain there are often multiple normative models which could be applied, and obeying one standard for rationality tends to lead to violations within other standards of rationality.

One method for dealing with the apparent dilemma is to propose that the mind is made up of a large number of specialized cognitive mechanisms, often referred to as "modules," which each embody their own internal standards of correct solutions within that particular domain. This is often identified as an evolutionary approach, although the same conclusion can be obtained via other routes (for example, via functional neuroanatomical evidence). One can similarly reach this conclusion by considering the implications of combinatorial explosion when trying to program problem solving machinery in artificial intelligence (i.e., the "frame problem"), which has been identified in philosophy as the problem of indeterminacy in inference (Quine, 1960; Dennett, 1978, 1984). It is also increasingly a commitment required to make sense of the precocious abilities of infants when tested using means such as the habituation paradigm (e.g., Wynn, 1992; Hirschfeld and Gelman, 1994; Wagner and Carey, 2003; McCrink and Wynn, 2004; Xu et al., 2004). The cognitive development field often refers to this situation as the existence of "constraints" in human infant mental abilities, reflecting the default assumption of a completely domain-general and content independent cognitive architecture. These constraints, however, are actually the *enablers* of specific cognitive abilities because the particular skills which they shunt infants into developing would not be able to emerge without the guidance of those constraints.

Although various people fret about this proposal being "massive modularity" (Samuels, 1998) or "modularity gone mad" (Fodor, 1987), it is the conclusion which the evidence impels us to accept. Besides indications that modularity is inevitable based on logical principles (Cosmides and Tooby, 1994; Callebaut and Rasskin-Gutman, 2005; Tooby and Cosmides, 2005; Carruthers, 2006; Ermer et al., 2007), computer simulations show that modularity is a consequence of neural organization under realistic conditions (Bullinaria, 2006; Clune et al., 2013), and actual physical and neurological structures point empirically to modular organization (Geary and Huffman, 2002; Cheverud et al., 2007). There is a functional carving of mental abilities, and it is a relatively fine-grained carving compared to what has generally been considered before (e.g., Inhelder and Piaget, 1958; Newell and Simon, 1972; Kahneman and Tversky, 1979; Johnson-Laird and Byrne, 1991). Furthermore, these lines of evidence do not commit anyone to impose some of the properties of modularity suggested by early ideas (Fodor, 1983). Functionally specialized cognitive modules are not required to be informationally encapulated, or to accept only highly local inputs, or to be reflexive and insensitive to contexts (see Barrett, 2005; Barrett and Kurzban, 2006 for in depth discussions of how and why these properties are not required elements). Some modules may in fact have these properties, but that does not mean that all modules must. In other words, the joints of nature may be carvable, but the lines are not necessarily "clean." Consider, by analogy, the various systems within the rest of the human body: respiration, digestion, circulation, etc. Many of these systems are intertwined, receiving inputs from each other and sending their outputs to other systems. Yet we still find it useful to separate these systems out for purposes of understanding and explaining them, and we can see the overall pattern of major functional adaptations embodied by these systems.

DO WE *REALLY* NEED TO CHANGE?

A skeptical reader might ask, "But, these are theoretical issues about the grand nature of the entire human mind (or all thinking); do I really need to change at the level my actual research? That is, what is my concern so long as I stick to my particular topic?" The response is that these issues of the grand nature of human thinking can and do percolate down to specific research topics. Attending to them opens up opportunities, and neglecting them creates problems.

Consider the area of human reasoning, a topic and field that is central to the idea of "thinking," and the most commonly used research tool in that field: Wason's selection task. The selection task was originally devised by Wason (1966) to evaluate if people can engage in logical falsification as part of, for example, scientific hypothesis evaluation. The task involves presenting a conditional rule (of the form, *If P, then Q*, where *P* and *Q* can be any content), usually some contextual information for the rule, and then four pictures of cards which are described as having relevant information printed on both sides of them. The visible sides of the four cards provide information about all four possible states relevant to the rule: *P*, *not-P*, *Q*, and *not-Q*. The task for participants is to indicate which of the cards need to be turned over for further information in order to evaluate the validity of the conditional rule. So, for instance, turning over the *P* card would reveal information (either *Q* or *not-Q*), and this is information which bears on the truth or falsity of the conditional rule.

The most traditional normative model for the selection task is first order conditional logic. Given a rule of the form "If *P*, then *Q*" (again, where *P* and *Q* are any content whatsoever), there are

logical conclusions that can be derived from additional pieces of information: if *P* is true, then *Q* is true; if *Q* is false (*not-Q*), then *P* is false (*not-P*). The cards which need to turned over for more information, according to formal logic, are the *P* and the *Not-Q* cards then, in order to assess if there are any violations of the rule. The general findings from Wason's original work and many subsequent studies is that people are notoriously poor at logical falsification such as this, even though it is computationally quite simple (e.g., trivially easy for a computer program to do; Newell et al., 1963).

Curiously, certain versions of the selection task eventually emerged on which people did quite well, even as they continued to perform poorly on the original version of the task. One of the most well known of these tasks which elicit good reasoning performance is the "drinking age problem" (Griggs and Cox, 1982), in which the conditional rule is *If a person is drinking beer, then they must be over 21"* (with the card options thus being: *Drinking beer, Drinking soda, 17 years old,* and *22 years old*). These content-based effects on reasoning bedeviled researchers and led some of them to seek out new theories and criteria by which to evaluate human reasoning abilities.

Human conditional reasoning does not follow the normative model of formal logic. But performance on Wason's selection task can be analyzed in terms of *other* normative models also (Elqayam and Evans, 2011). One can use deontic logic (conditional rules that regulate permissions and obligations) to evaluate correct versus incorrect responses. Cheng and Holyoak (1985), Holyoak and Cheng (1995) proposed that people induce pragmatic reasoning schemas, which closely parallel deontic logic principles, based on past experiences. Cosmides (1989), Cosmides and Tooby (1992) developed an explanation for selection task content effects based on evolved adaptations for reasoning about social contracts (conditional rules about reciprocal altruism, such as *If you take the benefit, then you must pay the associated cost*). One can alternatively use Bayesian reasoning or probability theory (Kirby, 1994; Oaksford and Chater, 1994, 1996, 2003; Oaksford et al., 1997) to evaluate correct versus incorrect responses in the selection task. In these models, correct responses are the selections which yield the highest expected information gain, whereas incorrect responses are those which yield little or no expected information gain. Finally, one can apply relevance theory (Girotto et al., 2001) to the selection task, proposing that the correct cards to select are the ones which are judged as most relevant to the current context.

This very cursory review of theories regarding human conditional reasoning illustrates a fundamental issue in terms of normative models in the study of human thinking. Most of these theoretical models of human reasoning aspire to be the one, best account of how human reasoning works. Researchers pit the models against each other, attempt to tally which model has the most support, best support, largest number of adherents, and so on. Which normative standard is the correct one? Which is correct at least in the case of human conditional reasoning? Apart from traditional disciplinary boundaries and preferences (or perhaps within-research laboratory traditions) there are no *a priori* justifications for these normative models. (And keep in mind that this illustration is just regarding conditional reasoning; it by no means exhausts the range of normative models for a realm as broad as "thinking.") The situation – the existence of content effects, the proliferation of normative-based models, the ongoing lack of consilience – points to there being no general normative models for all of human thought. One possible reaction is to largely abandon normativism (Elqayam and Evans, 2011). Another approach, which is taken here, is to recognize that there are different domains with different normative standards. Continuing to search for one normative model to rule all of human thought, or even all of human reasoning, is untenable and needs to change.

TIME SCALES AND RATIONALITY

There are several directions from which one can identify problems with the idea of general normative standards for rationality and human thought. Another aspect of this problem is illustrated by the tale of the village idiot:

> Once upon a time there was a village idiot who, when offered a choice between a dime and a nickel, would invariably choose the nickel. Everyone would laugh at the stupidity of the village idiot, and then go back to their chores until the next time they felt like a laugh. This went on for many years, during which the village idiot repeatedly and reliably chose to take a nickel over a dime. One day, a kind-hearted person tried to explain the situation to the village idiot. "Look, even though a nickel is larger than a dime, it is only worth half as much. So you should choose the dime." The idiot replied, "I know that. But if I choose the dime, people would stop offering me the choice between taking a nickel or a dime, wouldn't they? Who would be that stupid?"

The implicit normative standard that underlies this story is a standard economic utility model: people are rationally self-interested and should prefer a larger quantity of a desired item over a smaller quantity (Marshall, 1920). What the not-quite-such-an-idiot village idiot had done, however, was realize that there is always an implicit time scale when considering the utility of a sequence of events. A very small time scale, capturing just one event, can indicate one behavior as having the highest overall utility (a dime is better than a nickel). A different, longer, time scale, though (say, capturing at least three choices), can indicate a completely different behavior as having the highest overall utility.

The tale of the village idiot can be understood as parallel to the distinction between a one-shot prisoner's dilemma and a repeated prisoner's dilemma (Axelrod and Hamilton, 1981). The prisoner's dilemma is an economics game in which two people must decide whether to cooperate with the other person or defect against the other person. Mutual cooperation is rewarded, but not as much as defection when the other person cooperates (the "temptation payoff"). However, mutual defection does not pay as well as mutual cooperation, and cooperation when the other person defects yields a negative payoff (the "sucker's payoff"). A one-shot prisoner's dilemma has this payoff schedule and each person makes just one choice. In this one-shot version of the game the best strategy for each player is to defect, rather than cooperate, with the other player. As with the tale of the village idiot, this is based on the idea of utility maximization (in this case, maximization of the payoffs for each player) with a very small time scale of one move. Each player should go for the largest payoff (defecting), which also protects them from the worst outcome (being a sucker). If,

however, the prisoner's dilemma is played repeatedly between two players (also called an iterated game), there are strategies which are superior to constant defection in the longer time window. The most well known of these strategies is tit-for-tat, in which a player initially cooperates and then mirrors back whatever the previous choice was of the other player. Thus, two players can obtain the more modest (per play) reward of mutual cooperation rather than becoming stuck in mutual defection. These modest reward are *repeated* over the multiple rounds of the game. So, like the village idiot, each player accumulates multiple smaller payoffs which sum up to a much larger overall result than a single large payoff.

The effects of different time scale references also maps onto the idea of reciprocal altruism (Trivers, 1971) as a solution to the "problem of altruism" in biology. As evolutionary biologists considered the implications of evolutionary theory for the behavior of organisms, they realized that there seemed to be an overarching principle of complete self-interest: an individual should be focused intently on passing *their* genes into future generations and not at all interested – if anything, be antagonistic toward – other individuals managing to get their competing genes also into future generations. Yet in many cases animals did things which appeared to *help* other individuals, at a cost to themselves, which seemed to directly contradict the evolutionary theory implications. Hence, the "problem" of altruism. Along with the idea of kin-based altruism (Hamilton, 1964), a major explanation of these anomalous altruistic behaviors is the idea of reciprocal altruism (Trivers, 1971). The key insight for reciprocal altruism is that a single act of altruism (like cooperating in the prisoner's dilemma) can make sense if there is a reciprocal act of altruism with the roles reversed. So long as the value of the help experienced by each recipient is greater than their experienced cost of helping, there is a resulting net gain for both parties (known in economics as "gains in trade"). Once more, part of the key insight is to consider multiple, reciprocal behaviors between the two individuals – an expanded window of time rather than a thin slice.

DEEP RATIONALITY AND HUMAN THOUGHT

How can they say my life is not a success? Have I not for more than 60 years got enough to eat and escaped being eaten?

Smith (1931)

A resolution exists to this situation of arbitrarily conflicting normative models, many of which neglect the role of longer time scales, and it has been most fully and recently articulated by Griskevicius and Kenrick (2013), Kenrick and Griskevicius (2013), Kenrick et al. (2009), Kenrick et al. (2012). This resolution begins with a concept of "deep rationality," which presumes that rationality must be defined with respect to a very long time frame: the evolutionary selection pressures which shaped the human mind. There have been a multitude of different selection pressures and this insight, together with the idea of cognitive modularity, leads to the idea that there never was (and never will be) a single, proximate standard for normative rationality. Instead there are multiple motives which every person is balancing at any given time. In other words, to the extent that there is any overarching standard of rationality that designed our minds, it is not "don't screw up" but rather "survive and reproduce." This ultimate criteria, however, is not immediately useful beyond its ability to frame more specific problem solving domains (also see Buss, 1995; on top-level versus mid-level evolutionary theories). Rather than a general "survival and reproduction" criterion, this model presumes that there are different standards for a successful decision in different social problem domains such as: self-protection, status, affiliation, mate acquisition, mate retention, kin care, and disease avoidance. These domains provide fundamental motivational goals for people, but because there are several of them we can conceptualize our minds has having a number of different "subselves," each with different motives, different decision making processes, and even (from a more domain-general perspective) different cognitive biases.

Because different adaptive problems require different "rational" solutions, these solutions can only ever obey a local normative model which will inevitably break down once the topic under evaluation moves too far afield from the particular domain which constituted the evolutionary selection pressure and adaptive problem which created it (see also, Sperber, 1994 on the idea of proper domains for evolved mechanisms versus actual and cultural domains of application). Evolution designed many different cognitive programs, each embodying particular logics, designed to function well in particular contexts. In other words, the domain specificity of the cognitive mechanisms in the human mind implies that not only is there empirically no single normative standard of rationality which works across all of human cognition, but that there are good theoretical reasons why we should *expect* this to be the case.

This perspective belies many of the traditional criteria for normative models of rational though, such as obeying transitivity or the conjunction rule in probability; these are specifically applied as abstract, content-independent, and domain-general criteria. We should, in fact, be completely and utterly unsurprised that these types of criteria fail when they are applied to domains in which they do not correspond to the decision making adaptations evolution built within those domains. The fact that different theoretical models of conditional reasoning, as outlined above, each work particularly well within the context of particular reasoning contents should be alerting us to the fact that there is no one "human reasoning" normative model. Instead there are many cognitive mechanisms, each tailored to help us reason in an adaptive way about many different types of situations. It is even plausible that some limited abilities are included in this menagerie that enable general, abstract reasoning abilities when none of the specialized, evolutionarily-relevant contexts apply.

Or consider the prisoner's dilemma described earlier. Not only does using a different time scale change the nature of this dilemma, but specifying different players in the dilemma can change it as well. The classic prisoner's dilemma is played by two strangers (despite the allegorical "prisoners" being almost certainly friends). Strangers playing each other in the dilemma helps us to consider the situation more clearly in terms of domain-general, abstract rationality. But if, for instance, the prisoner's dilemma is played between biologically related individuals (kin), then issues of kinship and kin selection (Hamilton, 1964) come into play. The

payoffs within a prisoner's dilemma are fundamentally altered when the genetic fitness implications of playing with kin are factored into them: the points that a genetically related opponent obtains in a prisoner's dilemma are implicitly benefiting one's own biological fitness as well (due to the proportion of genes shared by virtue of common descent). For close kin, in fact, the dilemma actually resolves itself and there can be a mutually optimal equilibrium state (Kenrick et al., 2008, 2012). Strangers playing against each other in a prisoner's dilemma serves to simplify the situation, but it also is makes the situation less ecologically realistic; most of our real world interactions are with family, friends, and acquaintances.

So should "deep" evolutionary rationality serve as the definitive normative standard of behavior? Not necessarily. It is still critical to remember that human behavior has a foundation in cognitive adaptations, built by evolution over previous generations, and then further developed and filtered through our own experiences. A set of individual behaviors, within specific situations, can violate deep rationality, and violate it as a normative standard or as a descriptive standard. Being deeply rational is not the same as being omnipotent or omniscient. We are executers of cognitive programs (our evolved, mental adaptations). This means that there will be certain types of situations in which the cognitive programs produce "wrong" responses. One type of such situations is when there is an environmental mismatch: the responses which were shaped by many prior generations of evolution are no longer the best responses in our modern environment (e.g., our strong preferences for fats and sugars even when we already have enough; our general lack of desire for fiber in our diets even when we are in need of it). Another type of situation in which individual behaviors, based in deep rationality, can appear to be in violation of any normative or descriptive model is when there is a probabilistic outcome which is driving the selection pressure for that behavior (e.g., adolescent risk taking can appear to be irrational because it leads to some injuries and deaths, but at the same time if those behaviors produced an even larger social status and reputation benefit for the more successful risk takers then the overall behavioral tendency can be positively selected for nonetheless).

EXAMPLES

Having a child is surely the most beautifully irrational act that two people in love can commit.

Cosby (1987)

A couple of examples may help clarify the implications of taking a "deep rationality" perspective within a modular mind. The decision to have a child or not has been characterized as fundamentally sound (e.g., Holm, 2005), fundamentally unsound (e.g., Häyry, 2005), and even fundamentally impossible to evaluate (Paul, 2015 forthcoming). Certainly an economic cost/benefit analysis in modern environments does not support the position that having children is a rational choice. (The U.S. Department of Agriculture estimates that the cost of raising a child to the age of 17 is $269,520 (for families making over $70,200 per year.) On the other hand, a biological analysis would point out that reproduction is the most fundamental purpose of living organisms, and therefore any price is worth paying to have children. Somewhere in between these radical extremes are real people, who very often do opt to have children yet who also nearly always limit their reproductive rate to something significantly less than what they would theoretically accomplish if they devoted all their resources to having children. Brase and Brase (2012) found that both men and women have strong, emotional reactions (both positive and negative) to the prospect of having children, suggesting that there are countervailing forces at work in people's decisions about having children.

One compelling way to make sense of all these conflicting ideas and outcomes is to realize that desires to have children is but one of several different fundamental motives residing in people. We want to have children. But we also want to be safe (self-protection), we want to be respected (status), we want to be part of larger social groups (affiliation), we want to have and keep sexual partners (mate acquisition, mate retention), and we want to be healthy (disease avoidance).

Greed, for lack of a better word, is good. Greed is right, greed works.

[Gordon Gekko (Pressman and Stone, 1987)]

Now, a counterexample. The financial markets are perhaps the most elevated bastion of true and complete rationality. Adam Smith's "invisible hand" (Smith, 1776) rests on the idea of everyone acting in their rational economic self interests, and many people consider the Western financial markets to be a huge success of modern society. A closer look at the underlying foundations and assumptions of the modern financial market, however, can illustrate how its success arrives by stripping out all but one fundamental motive. The financial markets are not (or are minimally) interested in self-protection, status, affiliation, mate acquisition and retention, kinship, or disease avoidance. The financial markets are about money. With just one, clear motivating goal, it becomes possible to be completely rational in relation to the accomplishment of that goal. Critics of how the financial markets operate will often note, in various ways, this issue. Concerns include problems with the ethics of the financial markets, the effects of modern economic practices on human safety, security, or happiness. But these concerns are tangential to the central goal of the financial markets, so they form only externally imposed borders on behavior (e.g., through government regulations of disallowed actions).

If former presidential candidate Mitt Romney is correct that "corporations are people" (Rucker, 2011), what type of people are they? They are people who exist largely within the world of modern financial markets, and they therefore live lives that are single-mindedly about financial self-interest. Without all the other fundamental concerns that regular people have about their relationships with fellow humans, they quite possibly also qualify as psychopaths (Achbar et al., 2003). Before thinking that I am particularly anti-corporation, please note that it is also true that corporations are, by design and by law, exactly this way because we as a society have chosen to make them that way. Corporations cannot do anything other than act purely in their complete economic self interest. (Interesting things also, of course, occur due to the fact that corporations are often managed by regular humans who *do* recognize a multitude of other fundamental motives, and these corporation owners can elect to make decisions based on

their other motives, sometimes with the approval of shareholders and sometimes without).

CONCLUSION

This special topic in *Frontiers* (in which this article appears) describes the evolutionary approach to studying human thinking as *empirical normativism* in which human thinking is considered correct because it is the thinking which occurs (i.e., that there is no external evaluative standard). Such a view is described as a Panglossian framework, in which human thinking is considered *a priori* as being rational. This is unfair and incorrect.

First of all, to the extent that anyone actually exists who could be considered a Panglossian, this framework has never distinguished the evolutionary approach. This caricature of adaptationism is trafficked often by its critics and repeated by many who hear this criticism without realizing that it has been debunked repeatedly and by multiple, independent evaluations (e.g., Tooby and Cosmides, 1992; Borgia, 1994; Queller, 1995; Buss et al., 1998). Second, adaptationist models are not empirically driven, but rather based on evolutionary principles. The hypotheses (which are normative, in the sense of making predictions about what ought to happen, if the theory is correct) are based on careful consideration of evolutionary selection pressures, the constraints faced by a particular species, and existing evidence. Third, the appropriate issue is therefore not empirical normativism versus prescriptive normativism (which evaluates human thinking based on externally imposed criteria such as logic or probability theory), but rather how one should construct normative models of human thinking. Is it more useful to work with proximate models of normative rationality which proliferate under the traditional prescriptive normativism framework; models which becomes problematic as they struggle to accommodate *ad hoc*, competing domains of application? Or, is it better to work with higher level models of rationality, based on an evolutionary understanding of the central problems the mind has been sculpted to address?

An evolutionary framework, as outlined here, indicates that there are some normative standards which are useful for understanding the nature of human thinking, but that those standards are different from many of the normative standards proposed by prescriptive normativism. The search for a single normative model for all of human thinking is futile, because the multiple selection pressures which shaped the mind led to multiple cognitive mechanisms. A large-scale modularity of thinking processes is required, and in fact points toward useful ways to escape the multitudes of single-model theories which often stand in stalemates against each other. One specific version of this evolutionary modularity approach is the model of deep rationality (Kenrick et al., 2009, 2012), which specifies a set of fundamental motivational goals, each of which entails distinct patterns of reasoning and thinking (and which may be consistent, inconsistent, or orthogonal to each other).

ACKNOWLEDGMENTS

The author would like to thank Doug Kenrick and reviewers for helpful comments, and to Sandra Brase her ongoing advice and support.

REFERENCES

Achbar, M., Simpson, B., Achbar, M., and Abbott, J. (2003). *The Corporation [Motion Picture]*. Canada: Zeitgeist Films.

Axelrod, R., and Hamilton, W. D. (1981). The evolution of cooperation. *Science* 211, 1390–1396. doi: 10.1126/science.7466396

Barrett, H. C. (2005). Enzymatic computation and cognitive modularity. *Mind Lang.* 20, 259–287. doi: 10.1111/j.0268-1064.2005.00285.x

Barrett, H. C., and Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychol. Rev.* 113, 628–647. doi: 10.1037/0033-295X.113.3.628

Borgia, G. (1994). The scandals of San Marco. *Q. Rev. Biol.* 69, 373–375. doi: 10.1086/418652

Brase, G. L., and Brase, S. L. (2012). Emotional regulation of fertility decision making: what is the nature and structure of "baby fever?" *Emotion* 12, 1141–1154. doi: 10.1037/a0024954

Bullinaria, J. A. (2006). "Understanding the emergence of modularity in neural systems," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Vol. 31 (Vancouver, BC: Canada), 673–695.

Buss, D. M. (1995). The future of evolutionary psychology. *Psychol. Inq.* 6, 81–87. doi: 10.1207/s15327965pli0601_16

Buss, D. M., Haselton, M. G., Shackelford, T. K., Bleske, A. L., and Wakefield, J. C. (1998). Adaptations, exaptations, and spandrels. *Am. Psychol.* 53, 533–548. doi: 10.1037/0003-066X.53.5.533

Callebaut, W. E., and Rasskin-Gutman, D. E. (2005). *Modularity: Understanding the Development and Evolution of Natural Complex Systems.* Cambridge, MA: MIT Press.

Carruthers, P. (2006). "The case for massively modular models of mind," in *Contemporary Debates in Cognitive Science*, ed. R. J. Stainton (Malden, MA: Blackwell Publishing), 3–21.

Cheng, P. W., and Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cogn. Psychol.* 17, 391–416. doi: 10.1016/0010-0285(85)90014-3

Cheverud, J. M., Pavlicev, M., and Wagner, G. P. (2007). The road to modularity. *Nat. Rev. Genet.* 8, 921–931. doi: 10.1038/nrg2267

Clune, J., Mouret, J.-P., and Lipson, H. (2013). The evolutionary origins of modularity. *Proc. Biol. Sci.* 280:20122863. doi: 10.1098/rspb.2012.2863

Cosby, B. (1987). *Fatherhood.* New York: Berkley Books.

Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* 31, 187–276. doi: 10.1016/0010-0277(89)90023-1

Cosmides, L., and Tooby, J. (1992). "Cognitive adaptations for social exchange," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, eds J. H. Barkow, L. Cosmides, and J. Tooby (New York: Oxford University Press), 163–228.

Cosmides, L., and Tooby, J. (1994). "Origins of domain specificity: the evolution of functional organization," in *Mapping the Mind: Domain Specificity in Cognition and Culture: Mapping the Mind: Domain Specificity in Cognition and Culture*, eds L. Hirschfeld and S. Gelman (New York, NY: Cambridge University Press), 85–116.

Dennett, D. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology.* Montgomery, VT: Bradford Books.

Dennett, D. (1984). "Cognitive wheels: the frame problem in Artificial Intelligence," in *Minds, Machines and Evolution*, ed. C. Hookway (Cambridge: Cambridge University Press), 129–151.

Elqayam, S., and Evans, J. St. B. T. (2011). Subtracting 'ought' from 'is': descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Ermer, E., Cosmides, L., and Tooby, J. (2007). "Functional specialization and the adaptationist program," in *The Evolution of Mind: Fundamental Questions and Controversies*, eds S. W. Gangestad and J. A. Simpson (New York: Guildford Press), 153–160.

Fodor, J. A. (1983). *The Modularity of Mind.* Cambridge, MA: MIT Press.

Fodor, J. (1987). "Modules, frames, fridgeons, sleeping dogs, and the music of the spheres," in *Modularity in Knowledge Representation and Natural-Language Understanding*, ed. J. Garfield (Cambridge, MA: MIT Press), 26–36.

Geary, D. C., and Huffman, K. J. (2002). Brain and cognitive evolution: forms of modularity and functions of mind. *Psychol. Bull.* 128, 667–698. doi: 10.1037/0033-2909.128.5.667

Girotto, V., Kemmelmeier, M., Sperber, D., and van der Henst, J.-B. B. (2001). Inept reasoners or pragmatic virtuosos? Relevance and the deontic selection task. *Cognition* 81, B69–B76. doi: 10.1016/S0010-0277(01)00124-X

Griggs, R. A., and Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *Br. J. Psychol.* 73, 407–420. doi: 10.1111/j.2044-8295.1982.tb01823.x

Griskevicius, V., and Kenrick, D. T. (2013). Fundamental motives: how evolutionary needs influence consumer behavior. *J. Consum. Psychol.* 23, 372–386. doi: 10.1016/j.j.2013.03.003

Hamilton, W. D. (1964). The genetical evolution of social behavior 1 and 2. *J. Theor. Biol.* 73, 1–16, 17–57. doi: 10.1016/0022-5193(64)90039-6

Häyry, M. (2005). A rational cure for prereproductive stress syndrome. *J. Med. Ethics* 30, 377–378. doi: 10.1136/jme.2003.004424

Hirschfeld, L. A., and Gelman, S. A. (1994). *Mapping the Mind: Domain Specificity in Cognition and Culture.* New York: Cambridge University Press. doi: 10.1017/CBO9780511752902

Holm, S. (2005). Why it is not strongly irrational to have children. *J. Med. Ethics* 30:381. doi: 10.1136/jme.2003.004762

Holyoak, K. J., and Cheng, P. W. (1995). "Pragmatic reasoning about human voluntary action: evidence from Wason's selection task," in *Perspectives on Thinking and Reasoning: Essays in Honour of Peter Wason.* eds S. E. Newstead and J. S. B. T. Evans (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.), 67–89.

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence.* New York: Basic Books. doi: 10.1037/10034-000

Johnson-Laird, P. N., and Byrne, R. M. J. (1991). *Deduction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185

Kenrick, D. T., and Griskevicius, V. (2013). *The Rational Animal: How Evolution Made us Smarter than We Think.* New York: Basic Books.

Kenrick, D. T., Griskevicius, V., Sundie, J. M., Li, N. P., Li, Y. J., and Neuberg, S. L. (2009). Deep rationality: the evolutionary economics of decision making. *Soc. Cogn.* 27, 764–785. doi: 10.1521/soco.2009.27.5.764

Kenrick, D. T., Li, Y. J., White, A. E., and Neuberg, S. L. (2012). "Economic subselves: fundamental motives and deep rationality," in *Social Thinking and Interpersonal Behavior*, eds J. P. Forgas, K. Fiedler, and C. Sedikides (New York: Psychology Press), 23–43.

Kenrick, D. T., Sundie, J. M., and Kurzban, R. (2008). "Cooperation and conflict between kith, kin, and strangers: game theory by domains," in *Foundations of Evolutionary Psychology*, eds C. Crawford and D. Krebs (New York: Taylor and Francis Group/Lawrence Erlbaum Associates), 353–369.

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition* 51, 1–28. doi: 10.1016/0010-0277(94)90007-8

Marshall, A. (1920). *Principles of Economics. An Introductory Volume,* 8th Edn. London: Macmillan.

McCrink, K., and Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychol. Sci.* 15, 776–781. doi: 10.1111/j.0956-7976.2004.00755.x

Newell, A., Shaw, J. C., and Simon, H. A. (1963). "Empirical explorations with the logic theory machine: a case study in heuristics," in *Computers and Thought*, eds E. A. Feigenbaum and J. Feldman (New York, NY: McGraw Hill), 109–133.

Newell, A., and Simon, H. A. (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 608–631. doi: 10.1037/0033-295X.101.4.608

Oaksford, M., and Chater, N. (1996). Rational explanation of the selection task. *Psychol. Rev.* 103, 381–391. doi: 10.1037/0033-295X.103.2.381

Oaksford, M., and Chater, N. (2003). Optimal data selection: revision, review, and reevaluation. *Psychon. Bull. Rev.* 10, 289–318. doi: 10.3758/BF03196492

Oaksford, M., Chater, N., Grainger, B., and Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 441–458. doi: 10.1037/0278-7393.23.2.441

Paul, L. A. (2015, forthcoming). What you can't expect when you're expecting. *Res. Philosophica* 92, 1–23. doi: 10.11612/resphil.2015.92.2.1

Pressman, E. R., and Stone, O. (1987). *Wall Street [Motion picture].* United States: 20th Century Fox.

Queller, D. C. (1995). The spandrels of St. Marx and the Panglossian paradox: a critique of a rhetorical programme. *Q. Rev. Biol.* 70, 485–490. doi: 10.1086/419174

Quine, W. V. (1960). *Word and Object.* Cambridge, MA: MIT Press.

Rucker, P. (2011). *Mitt Romney Says 'Corporations are People' at Iowa State Fair.* Available at: http://articles.washingtonpost.com/2011-08-11/politics/

Russell, B. (1950). *Unpopular Essays.* London: George Allen and Unwin.

Samuels, R. (1998). Evolutionary psychology and the massive modularity hypothesis. *Br. J. Philos. Sci.* 49, 575–602. doi: 10.1093/bjps/49.4.575

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations. London: W. Strahan.*

Smith, L. P. (1931). *Afterthoughts.* London: Constable & Company, Inc.

Sperber, D. (1994). "The modularity of thought and the epidemiology of representations," in *Mapping the Mind: Domain Specificity in Cognition and Culture,* eds L. A. Hirschfeld and S. A. Gelman (New York, NY: Cambridge University Press), 39–67.

Tooby, J., and Cosmides, L. (1992). "The Psychological Foundations of Culture," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, eds J. H. Barkow, L. Cosmides, and J. Tooby (Oxford, England: Oxford University Press.), 19–136.

Tooby, J., and Cosmides, L. (2005). "Conceptual foundations of evolutionary psychology," in *The Handbook of Evolutionary Psychology*, ed. D. M. Buss (Hoboken, NJ: John Wiley & Sons Inc.), 5–67.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35–57. doi: 10.1086/406755

Wagner, L., and Carey, S. (2003). Individuation of objects and events: a developmental study. *Cognition* 90, 163–191. doi: 10.1016/S0010-0277(03)00143-4

Wason, P. C. (1966). "Reasoning," in: *New Horizons in Psychology,* ed. B. M. Foss (Harmondsworth: Penguin), 135–151.

Wynn, K. (1992). Evidence against empiricist accounts of the origins of numerical knowledge. *Mind Lang.* 7, 315–332. doi: 10.1111/j.1468-0017.1992.tb00306.x

Xu, F., Carey, S., and Quint, N. (2004). The emergence of kind-based object individuation in infancy. *Cogn. Psychol.* 49, 155–190. doi: 10.1016/j.cogpsych.2004.01.001

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 November 2013; accepted: 24 April 2014; published online: 15 May 2014.
Citation: Brase GL (2014) The nature of thinking, shallow and deep. Front. Psychol. ***5****:435. doi: 10.3389/fpsyg.2014.00435*
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 20 August 2014
doi: 10.3389/fpsyg.2014.00901

Reason and less

Vinod Goel [1,2] *

[1] Department of Psychology, York University, Toronto, ON, Canada
[2] Istituto di Ricovero e Cura a Carattere Scientifico-Fondazione Ospedale San Camillo, Venice, Italy

Edited by:
David E Over, Durham University, UK

Reviewed by:
Ira Andrew Noveck, Centre Nationale de la Recherche Scientifique, France
Gary L. Brase, Kansas State University, USA

***Correspondence:**
Vinod Goel, Department of Psychology, York University, Toronto, ON M3J 1P3, Canada
e-mail: vgoel@yorku.ca

We consider ourselves to be rational beings. We feel that our choices, decisions, and actions are selected from a flexible array of possibilities, based upon reasons. When we vote for a political candidate, it is because they share our views on certain critical issues. When we hire an individual for a job, it is because they are the best qualified. However, if this is true, why does an analysis of the direction of shift in the timbre of the voice of political candidates during an exchange or debate, predict the winner of American presidential elections? Why is it that while only 3% of the American population consists of white men over 6′4″ tall, 30% of the CEOs of Fortune 500 companies are white men over 6′4″ tall? These are examples of "instinctual biases" affecting or modulating rational thought processes. I argue that existing theories of reasoning cannot substantively accommodate these ubiquitous, real-world phenomena. Failure to recognize and incorporate these types of phenomena into the study of human reasoning results in a distorted understanding of rationality. The goal of this article is to draw attention to these types of phenomena and propose an "adulterated rationality" account of reasoning as a first step in trying to explain them.

Keywords: rationality, reasoning, Decision Making, evolutionary psychology, instincts, biases

Nature, Mr. Allnutt, is what we were put on this world to rise above.
–Katharine Hepburn to Humphrey Bogart in *The African Queen*

INTRODUCTION AND BACKGROUND

The conception of man as a rational, thinking being, permeates Western thought from (at least) Aristotle to present times. Our behavior is explained by postulating beliefs and desires, and a principle of "rationality," that guides our pursuit of the latter in the context of the former. My use of the term "rationality" is derived from the philosophical literature, meaning roughly, deliberate reason. (It does not imply a commitment to any normative standard, as is the case in the psychology literature.) On this account rationality is goal directed behavior. It is simply a means to an end and is ascribed to individual agents. It is a deliberate choice or action that moves an organism closer to its goal in a manner consistent with its knowledge and beliefs. A rational choice is not simply a selection, it is a selection for a reason (Bermudez, 2002). Perhaps the most significant feature of a rational system is the existence of a "gap" between the stimulus and the response (Cassirer, 1944). The stimulus or antecedent condition is never causally sufficient to determine any specific choice or action. I will use the term "reason-based" to refer to this general notion of rationality. Reason-based choice is often contrasted with instinctual or tropistic behavior, where there is no such gap and the antecedent conditions are causally sufficient for a course of action (Bermudez, 2002).

Consider the following example: when a young male lion chases away an older male and takes over a pride, he proceeds to kill any cubs the females may be nursing. How should we explain this behavior? Does the lion sit down and reason thus: "these cubs do not perpetuate my genes. They will require the expenditure of considerable resources to feed and defend. Providing these resources to perpetuate someone else's genes does not make evolutionary sense. However, if I kill these cubs (which I surely can, without harm to myself), the females will stop lactating and come into heat again. I can then impregnate them with my sperm and then they will bear my offspring. Then the resources of the pride can be used to propagate my genes rather than someone else's. Therefore, it is reasonable to kill these cubs."

If the lion did deliberate in this way, we would be justified in saying his behavior, however, cruel, was rational. If he reasoned thus, and did not kill the cubs, his behavior would be irrational. But most of us do not believe that the lion has the ability to reason in this manner. Most of us do not believe that the lion chooses actions from a vast array of possible alternatives *for reasons*. His behavior is compelled, in the context of particular environmental and developmental factors. Therefore, applying the label of "rationality" (given the above definition) to explain this behavior is both unnecessary and incorrect. The lion's behavior is certainly adaptive, but it is not rational or reason-based. It is explained by appeals to instinctual or tropistic mechanisms (such as parental investment parasitization prevention) that are triggered by causal interactions between the maturation of specific internal structures and environmental cues. Once the mechanisms are triggered, they lead to a particular course of action.

Now consider the case of a man who partners with a woman with young children from a previous partner. The man does not typically kill the children, though it may be, arguably, adaptive to do so. Why not? Presumably because he's making a conscious choice from a wide range of possibilities. He is not driven to an inevitable action. He could choose not to get involved in this relationship and find a woman without children. Perhaps he loves the children. Perhaps he finds the woman so desirable that the opportunity to have his own children with her is worth the price

of expending some resources to raise her children from a previous partner. Perhaps his overtures to women without children have been unsuccessful. Whatever the reason, he is making a conscious, rational/reason-based choice.

However, if this appeal to reason is adequate to explain the behavior of the man, why is it the case that instances of child abuse/mistreatment are much higher in the case of stepfathers (and stepmothers) than biological fathers and mothers (Daly and Wilson, 2005)? Why is it the case that, despite our convictions that we vote for political candidates because they share our views on certain critical issues, that a simple analysis of the direction of shift in the timbre of the voice of candidates during an exchange or debate, predicts the winner of American presidential elections (Gregory and Gallagher, 2002)? Why is it the case that, despite our beliefs that we hire individuals for jobs because they are the best qualified, 30% of the CEOs of Fortune 500 companies are white men over 6′4″ tall, even though they represent only 3% of the American population (Rauch, 1995)?

These are all examples of what we might call "instinctual biases" affecting our reasoning and decision-making processes (Buss, 2005). They are genuine, ubiquitous phenomena. But they are not phenomena typically studied by cognitive psychologists interested in human reasoning. Our current research programs either (1) ignore these types of "biases" (Rips, 1994; Johnson-Laird, 2006); (2) assume that they are cut from the same cloth as the conceptual biases in the Linda problem (see below) and can be explained in the same fashion (Evans and Over, 1996; Stanovich, 2004); (3) focus on instinctual biases, but assume that is all there is to human reasoning (Cosmides and Tooby, 1994b; Duchaine et al., 2001) or (4) consider them to be social biases built on top of the cognitive engine and as such do not influence the operation of that engine (Berry, 2007).

I want to suggest that ignoring these phenomena excludes much of what is interesting about human reasoning from our research programs, and may, in fact, result in distorted theories of human reasoning based upon incomplete data sets. Furthermore, if evolutionary psychologists are correct, the effect of biological markers such as dominance cues, facial attractiveness cues, waist to hip ratios (in women), shoulder to waist ratios (in men), etc. are not socially construed phenomena, but apply universally (Buss, 2005). The two theories of reasoning best situated to account for these phenomena are massive modularity theory (Cosmides and Tooby, 1994b) and dual mechanism theories (Sloman, 1996; Evans, 2003; Stanovich, 2004). I argue that neither of these accounts can adequately accommodate the phenomena and propose a banal adulterated rationality account.

CONCEPTUAL SPACE OF THEORIES OF HUMAN REASONING

INFORMATION PROCESSING THEORY

Classical information processing theory holds that the cognitive system is a general purpose information processing system, perhaps with some specialized modules, for language and perceptual processes (Fodor, 1983; Pylyshyn, 1984; Newell, 1990). In the context of this framework there are several accounts of reasoning such as mental model theory (Johnson-Laird, 2006) and mental logic theory (Braine, 1978; Rips, 1994). While there are significant differences between these two theories, in terms of the nature of the representations and computations employed during logical reasoning, both postulate a mechanism that operates within the rules of formal logic. These theories generally do not try to explain the phenomenon of "instinctual biases" of interest here.

MASSIVE MODULARITY

There is a research program that explicitly sets out to account for instinctual biases. Indeed, 20 years ago Cosmides and Tooby (1994a) exhorted cognitive scientists not to be blind to the effect of instincts ("instinct blindness"). With respect to reasoning, they have worked largely with the Wason card selection task (Cosmides, 1989; Fiddick et al., 2000). The Wason card selection task is a disguised form of conditional inference. Four cards, corresponding to the four forms of the conditional (modus ponens, modus tollens, denying the antecedent, and affirming the consequent) are placed in front of the subject, along with the conditional rule. The basic result is that switching from a rule with arbitrary content (e.g., "if the letter on one side of the card is a vowel, then the number on the other side must be even") to a rule that embodies the structure of some social contract (e.g., "if someone is drinking beer, then they must be over 18 years of age"), increases performance accuracy from the order of 6% to the order of 80% (Wason and Shapiro, 1971; Cox and Griggs, 1982). The explanation is that this dramatic shift in performance is the result of the triggering of a "cheater detection" module.

On the massive modularity account the mind is not a general purpose information processing system, but rather consists of 1000s of special-purpose modules selected for the adaptive advantage they conferred upon our Pleistocene ancestors in solving problems specific to their environment, such as selecting mates, leaders, and detecting cheaters (Pinker, 1997; Duchaine et al., 2001). The modules are causally triggered by specific environmental cues. In previous times we would have called these modules instincts. Today we might liken them to the apps on our smartphones (Kurzban, 2012). Like apps they work relatively independently, though they may have access to information generated by other specific apps. For example, the app that I use to monitor my walks has access to information from the GPS and the system clock. It does not have (nor requires) access to the output of the apps that I use to listen to audiobooks or track flight arrivals. One can of course imagine a greater degree of interdependence and interaction among modules, but the main point is that there is no general-purpose reasoning system that controls the selection and triggering of individual modules. The selection and triggering are determined by direct causal links to specific environmental cues. On a strong version of the account, it is claimed that our notion of rationality (or general purpose reasoning) is illusory. What we regard as "general purpose reasoning" is just the functioning of numerous instinctual modules (Cosmides and Tooby, 1994b).

Several authors offer compelling critiques of the massive modularity account (Fodor, 2001; Over, 2002). I believe its greatest strength is that it offers a potential solution to the intractable problem of induction (or the frame problem) that plagues cognitive psychology, albeit at the price of a tight causal coupling between specific environmental cues and triggering of specific modules. Its greatest weakness is that it cannot explain how we can

send a man to the moon and predict and investigate the existence of the Higgs boson (because presumably nothing in the Pleistocene environment of our ancestors would have selected for these abilities).

But for present purposes, I will limit my concerns about massive modularity to its ability to explain the specific examples of reasoning/decision-making with which I began. It is not clear that the above examples can be explained *just* in terms of activation of specific instinctive modules. One reason to doubt the ability of the massive modularity model to explain the phenomenon of interest is to note the differences in the response patterns, and the ability of participants to reflect upon and justify their responses, in the case of our examples and the Wason card selection task.

In the case of the Wason card selection task, a shift in content to a rule breaking scenario results in a shift in accuracy approaching ceiling level, and one can plausibly argue that the shift in content triggers something like a cheater detection module. However, this does not seem to be the case for the examples in the introduction. For example, while instances of child abuse by stepparents are significantly greater than for biological parents, they do not approach 80% (Daly and Wilson, 2005). We will not typically vote for a leader who intends to adversely affect our lives, despite the presence of dominance cues. In hiring a doctor we would not typically choose a tall, handsome, athletic man without a medical degree over a short, hunchbacked, pudgy man with a medical degree. Therefore these phenomena cannot be explained just in terms of an appeal to instinctual modules (as they can in the case of the lion, and perhaps even the Wason card selection task). Such phenomena call for *a blended response* between instincts/modules and some general purpose reasoning system.

There is also a discrepancy between the response/behavior and the reason/explanation offered for the behavior. In the case of the Wason card selection task, participants can typically articulate why they chose particular cards (in the familiar content or cheater detection version). In the case where we are evaluating two potential employees (or grad students) with similar views and qualifications, if one exhibits high attractiveness cues, while the other does not, we will often choose the attractive individual, but when questioned, our explanation will not implicate these cues. It will be in terms of the qualifications of the candidates, even though there may be no material differences in these factors (Dipboye et al., 1975; Langlois et al., 2000; Hosoda et al., 2003). We will not be consciously aware of the effect of the attractiveness cues on our reasoning/decision-making behavior. This again suggests that there are at least two processes at work here, a conscious general-purpose reasoning system that evaluates the qualifications of the two candidates, and unconscious instinctual biases that modulate the operation of the former system.

It may be tempting to draw parallels between our inability to report on the causal efficaciousness of instinctual biases and the confabulation that split brain patients engage in when the verbal left hemisphere is unaware of the choices made by the right hemisphere. While there are some similarities, the dissimilarities may be greater. Consider the following famous experiment (Gazzaniga, 1998): a split brain patient was presented with a picture of a winter scene projected to the right hemisphere (left visual field) and a picture of a chicken claw projected to the left hemisphere (right visual field). The patient must then select two related pictures, one picture with each hand, from an array of other pictures. The patient's left hand points to a shovel (because the right-hemisphere, controlling that hand has seen a snow-covered winter scene) and the right-hand points to a chicken (because the left hemisphere, controlling that hand, has seen the chicken claw). When the patient is asked to explain why his left hand (guided by the right hemisphere) is pointing to the shovel, the left hemisphere (dominant for language) has no access to the information about the winter scene seen by the right hemisphere. But instead of responding "I don't know" the patient responds by noting that the shovel is required to clean the chicken coop.

The similarity lies in the fact that in both cases, the verbal explanation for the behavior cannot causally account for the behavior. The dissimilarity is that the explanation offered by the left hemisphere of the split brain patient is a complete *post hoc* confabulation. It simply is not relevant to explaining the behavior. In the case of instinctual biases, we are not "confabulating" in the same sense because the conscious explanation that we offer (e.g., "this applicant has a degree from University of Waterloo") is usually causally relevant. It cannot explain the complete pattern of the data, but it may be a relevant part of the causal story.

DUAL MECHANISM THEORIES

There is a research program that acknowledges the necessity of a general-purpose reasoning system and also explicitly sets out to account for various reasoning biases. These dual systems accounts of reasoning contrast heuristic/intuitive (System 1) processes with formal (System 2) processes (Sloman, 1996; Evans, 2003; Stanovich, 2004). This is becoming a widely accepted distinction and seems to have an underlying neuropsychological basis (Goel and Dolan, 2003; Goel, 2007). The critical feature of this paradigm is that while there is a logical/formal response to the task, in some conditions it is inhibited and bypassed by subjects' *background knowledge and beliefs*. An example is provided by the famous Linda Problem (Tversky and Kahneman, 1983):

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*
> *Which statement is most likely?*
>
> *(a) Linda is a bank teller and active in the feminist movement*
> *(b) Linda is a bank teller*

The effect is that many intelligent individuals will choose the conjunction (a) as more likely than one of the conjuncts (b). Their rationale is that the conjunct (b) by itself does not seem sufficient for someone with Linda's background. The conjunction (a) in addition contains a conjunct that seems more appropriate given the background description of Linda. The usual explanation for the "irrational" response is that overall (a) is more "representative" of Linda than (b) (even though a conjunction cannot be more likely than either conjunct)[1]. This has led to a distinction between

[1]The phenomenon disappears if one of the conjuncts is "Linda is active in the feminist movement." or if the conjunction is "Linda is a bank teller and is 43 years old."

formal processes and heuristic/intuitive processes (Evans, 2003; Stanovich, 2004).

This is a genuine, important psychological phenomenon. Some dual mechanism theorists have argued that these heuristic responses represent primitive, low-level instinctual biases that we share with pigeons and rats (Evans and Over, 1996). This is simply a mistake. The bias exhibited in the Linda problem is a very high level, conceptual bias based upon language and our knowledge of the world. Note that while heuristic responses may be considered irrational, in sense of violating normative logic, [though this is a moot point (Politzer and Noveck, 1991; Gigerenzer, 2007; Goel, 2008)], both responses are clearly reason-based, as I am using the term here. There are sensible psychosocial expectancy reasons for why subjects choose the so-called "irrational" response. If the logical inconsistency of their response is pointed out to subjects, they can quickly give the logically correct response and offer justifications for the initial heuristic response (Sloman, 1996). My conjecture is that instinctual biases are drawn from a very different well than the conceptual biases exhibited in the Linda problem. If this is the case, there is no reason to believe that the theory can account for the types of reasoning phenomenon of interest here.

One response to this objection is to note that System 1 is a heterogeneous collection of everything from reflex arcs to conceptual biases (Stanovich, 2004). In this case, the instinctual biases I am trying to bring attention to would fall into System 1 (as would many reason-based processes). Even though it has been argued elsewhere (Goel, 2008), that the differences in the underlying causal mechanisms of such a heterogeneous collection of System 1 processes makes the category uninteresting for theory building, the distinction is considered useful because System 1 processes are said to share behavioral similarities in outputs in terms of speed and automaticity of responses (Stanovich, 2004). Here I want to suggest that the behavioral patterns are very different in the case of conceptual biases and instinctual biases. The argument here is similar to that offered above for the massive modularity account.

The whole point of dual mechanism accounts is that the processing goes through one of the two systems. The responses are *either* "rational" *or* "heuristic." This model works well for the type of phenomena the theory was intended to explain, such as content effects in syllogisms (Evans, 2003) and the Linda problem (Tversky and Kahneman, 1983). In each of these cases the conceptual biases result in a dramatic shift in subject responses, so perhaps one can argue that the bias results in the queuing of a different system. Furthermore, individuals are consciously aware of and can articulate the reasons for the "non-rational" response (Sloman, 1996).

In the case of the instinctual biases of interest here, there is neither a dramatic shift in behavior such that 90% of participants are responding in one way or the other, nor an awareness of the reasons for the behavioral shift. Both of these points have been illustrated with examples in the above discussion of massive modularity. As in the case of massive modularity, the dual mechanism accounts work best when there is a dramatic shift in performance (as in the Linda Problem) but will require some sort of modulation/interaction account where the response shift is graded and less pronounced, and subjects are unable to fully articulate causally efficacious reasons for their response/choice.

ADULTERATED RATIONALITY ACCOUNT

I think the key missing feature in the above accounts of human reasoning is the recognition of the modulation of rational choice by instinctual biases. Any theory that is going to do justice to human reasoning must acknowledge both a rational system and a host of instinctual systems or biases. It must also acknowledge that these systems interact, to varying degrees, in human reasoning and decision-making. Human choices cannot be explained by postulating a single type of system, whether it be instinctual modules or a general-purpose reasoning system.

I am proposing a banal model whereby the rational engine has evolved on top of instinctual/tropistic mechanisms. The nature of these instinctive mechanisms can perhaps be understood along the lines of the "automatic appetitive impulsive processes" postulated in the addiction literature (Gladwin et al., 2011; Wiers et al., 2013). The instinctual biases of interest here need not be "appetitive" processes, but they are automatic, impulsive, non-cognitive processes that manifest individual differences, and modulate and in turn are modulated by, top-down (reason-based) executive control processes. Thus functioning of the rational engine is modulated or adulterated by these processes to varying degrees, depending on the nature of the tasks, and individual differences. For example, the rational engine would be more affected by instinctual biases in the case of mate selection than in calculating the launch trajectory of a satellite to orbit Mars. I view this process of modulation or adulteration as one of bending and warping the architecture of the reason-based system such that certain possibilities are facilitated, hindered, or even blocked. I propose to call this the "adulterated rationality" account of reasoning.

The system is set up in such a way that the unconscious bottom–up instinctual biases or modules are triggered by task specific cues in the environment (along the lines postulated by the massive modularity account), however, rather than being the sole determinants of behavior, these biases pass through a conscious top-down reason-based system, resulting in a response that is a blended product of the two systems. Individual differences in the strength of specific bottom-up, non-cognitive, instinctual biases, and the strength of top-down cognitive, reason-based processes and strategies, along with the nature of the reasoning task, will affect the ratio of the mixture.

For example, consider the discriminative parental solicitude effect with which we began. Parental investment is a valuable resource that can be parasitized by non-relatives (Daly and Wilson, 1994). It is suggested that we all have mandatory, automatic, innate mechanisms for countering parental investment parasitism (Daly and Wilson, 1994). These mechanisms must be suppressed in the case of stepfathers (and stepmothers) where they make the conscious decision to accept a mate with children from another partner. The majority of stepfathers and stepmothers are able to bond (to some extent) with their new mate's existing offspring, but there is considerable individual variability. The standard explanation for failure would implicate top down inhibition processes (i.e., "they didn't try hard enough"). But an equally likely possibility is variability in the strength of the mandatory impulsive, non-cognitive, bottom-up processes. If these instinctual systems are exceptionally

strong in certain individuals, then an equivalent exertion of top-down processes will not result in the same effect. This raises some interesting psychological, biological, ethical, and legal issues.

CONCLUSION

My goal here has been to draw attention to an ubiquitous, but neglected phenomenon which affects our rational behavior: the modulation of conscious rational choice by unconscious instinctual biases. Much of the study of human rationality within cognitive psychology has focused on logical form. It is time to look beyond logical form. Recent studies directed at the role of emotions in logical reasoning are beginning to do this (Blanchette, 2006; Goel and Vartanian, 2010). However encouraging, this is not sufficient. We need to cast a much broader net and incorporate the type of phenomena identified here. Failure to do so will result in incomplete and distorted theories of reasoning. Broadening the research program means developing experimental paradigms to study the role of instinctual biases on decision-making and using these data to inform cognitive theory. I believe that incorporating these data will point us toward something like an adulterated rationality account of reasoning.

Furthermore, cognitive psychology has emphasized the importance of top-down cognitive inhibitory processes in understanding human behavior. We know something about the neuropsychology of these processes (Shallice and Cooper, 2011). However, the adulterated rationality model, in identifying the importance of bottom-up, non-cognitive, instinctual processes, and recognizing individual differences, suggests that this focus is only half of the story. Deviation of behavior from expected norms may not simply be a function of failure of top-down control, but individual differences in the strength of the bottom-up processes. If this is the case, it would have important consequences for our legal and social norms and expectations.

Thus in summary, I am drawing attention to ubiquitous, real-world, reasoning paradigms where rational choice is modulated by instinctual biases. I argue that existing models of logical reasoning cannot adequately accommodate these phenomena and propose an adulterated rationality account of reasoning. The ubiquitousness of the phenomena call for data collection, model fitting and exploration of consequences for social and legal norms.

ACKNOWLEDGMENTS

This work was funded, in part, by an NSERC grant and Wellcome Trust Grant (089233) to Vinod Goel.

REFERENCES

Bermudez, J. L. (2002). "Rationality and psychological explanation without language," in *Reason and Nature: Essays in the Theory of Rationality,* eds J. L. Bermudez and A. Millar (New York: Oxford University Press), 233–264.

Berry, B. (2007). *Beauty Bias: Discrimination and Social Power.* London: Greenwood Publishing Group, Inc. doi: 10.1037/0033-295X.85.1.1

Blanchette, I. (2006). The effect of emotion on interpretation and logic in a conditional reasoning task. *Mem. Cognit.* 34, 1112–1125. doi: 10.3758/BF03193257

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychol. Rev.* 85, 1–21. doi: 10.1037/0033-295X.85.1.1

Buss, D. M. (2005). *The Handbook of Evolutionary Psychology*, 1st Edn. Hoboken, NJ: Wiley.

Cassirer, E. (1944). *An Essay On Man: An Introduction to a Philosophy of Human Culture.* New Haven: Yale University Press.

Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task [see comments]. *Cognition* 31, 187–276. doi: 10.1016/0010-0277(89)90023-1

Cosmides, L., and Tooby, J. (1994a). Beyond intuition and instinct blindness: toward an evolutionarily rigorous cognitive science. *Cognition* 50, 41–77. doi: 10.1016/0010-0277(94)90020-5

Cosmides, L., and Tooby, J. (1994b). "Origins of domain specificity: the evolution of functional organization," in *Mapping the Mind: Domain Specificity in Cognition and Culture,* eds L. Hirschfeld and S. Gelman (NewYork: Cambridge University Press).

Cox, J. R., and Griggs, R. A. (1982). The effects of experience on performance in Wason's selection task. *Mem. Cognit.* 10, 496–502. doi: 10.3758/BF03197653

Daly, M., and Wilson, M. (2005). The "Cinderella effect" is no fairy tale. *Trends Cogn. Sci.* 9, 507–508; author reply 508–510. doi: 10.1016/j.tics.2005.09.007

Daly, M., and Wilson, M. I. (1994). Some differential attributes of lethal assaults on small children by stepfathers versus genetic fathers. *Ethol. Sociobiol.* 15, 207–217. doi: 10.1016/0162-3095(94)90014-0

Dipboye, R. L., Fromkin, H. L., and Wiback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. *J. Appl. Psychol.* 60, 39–43. doi: 10.1037/h0076352

Duchaine, B., Cosmides, L., and Tooby, J. (2001). Evolutionary psychology and the brain. *Curr. Opin. Neurobiol.* 11, 225–230. doi: 10.1016/S0959-4388(00)00201-4

Evans, J. (2003). In two minds: dual-process accounts of reasoning. *Trends Cogn. Sci.* 7, 454–459. doi: 10.1016/j.tics.2003.08.012

Evans, J., and Over, D. E. (1996). *Rationality and Reasoning.* New York, NY: Psychology Press.

Fiddick, L., Cosmides, L., and Tooby, J. (2000). No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. *Cognition* 77, 1–79. doi: 10.1016/S0010-0277(00)00085-8

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology.* Cambridge, MA: MIT Press.

Fodor, J. A. (2001). *The Mind Doesn't Work That Way.* Cambridge, MA: MIT Press.

Gazzaniga, M. S. (1998). *The Mind's Past.* Berkeley: University of California Press.

Gigerenzer, G. (2007). *Gut Feelings.* New York: Viking.

Gladwin, T. E., Figner, B., Crone, E. A., and Wiers, R. W. (2011). Addiction, adolescence, and the integration of control and motivation. *Dev. Cogn. Neurosci.* 1, 364–376. doi: 10.1016/j.dcn.2011.06.008

Goel, V. (2007). Anatomy of deductive reasoning. *Trends Cogn. Sci.* 11, 435–441. doi: 10.1016/j.tics.2007.09.003

Goel, V. (2008). "Fractionating the system of deductive reasoning," in *The Neural Correlates of Thinking*, eds E. Pppel, B. Gulyas, and E. Kraft (New York: Springer Science).

Goel, V., and Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition* 87, B11–B22. doi: 10.1016/S0010-0277(02)00185-3

Goel, V., and Vartanian, O. (2010). Negative emotions can attenuate the influence of beliefs on logical reasoning. *Cogn. Emot.* 25, 121–131. doi: 10.1080/02699931003593942

Gregory, S., and Gallagher, T. (2002). Spectral analysis of candidates' nonverbal vocal communication: predicting US presidential election. *Soc. Psychol. Q.* 65, 298–308. doi: 10.2307/3090125

Hosoda, M., Stone-Romero, E. F., and Coats, G. (2003). The effects of physical attractiveness on job-related outcomes: a meta-analysis of experimental studies. *Pers. Psychol.* 56, 431–462. doi: 10.1111/j.1744-6570.2003.tb00157.x

Johnson-Laird, P. (2006). *How We Reason.* Oxford: Oxford University Press.

Kurzban, R. (2012). *Why Everyone (Else) Is a Hypocrite: Evolution and the Modular Mind.* Princeton: Princeton University Press.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., and Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychol. Bull.* 126, 390–423. doi: 10.1037/0033-2909.126.3.390

Newell, A. (1990). *Unified Theories of Cognition.* Cambridge, MA: Harvard University Press.

Over, D. (2002). "The rationality of evolutionary psychology," in *Reason and Nature: Essays in the Theory of Rationality,* eds J. L. Bermudez and A. Millar (New York: Oxford University Press), 187–207.

Pinker, S. (1997). *How the Mind Works.* New York: Norton & Co.

Politzer, G., and Noveck, I. A. (1991). Are conjunction rule violations the result of conversational rule violations? *J. Psycholinguist. Res.* 20, 83–103. doi: 10.1007/BF01067877

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.

Rauch, J. (1995). Short guys finish last. *The Economist.* Available at: http://www.jonathanrauch.com/jrauch_articles/2004/08/short_guys_fini.html (accessed December 23, 1995).

Rips, L. J. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA: MIT Press.

Shallice, T., and Cooper, R. (2011). *The Organization of Mind*. Oxford: Oxford University Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 119, 3–22. doi: 10.1037/0033-2909.119.1.3

Stanovich, K. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago: University of Chicago Press. doi: 10.7208/chicago/9780226771199.001.0001

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Wason, P. C., and Shapiro, D. A. (1971). Natural and contrived experience in a reasoning problem. *Q. J. Exp. Psychol.* 23, 63–71. doi: 10.1080/00335557143000068

Wiers, R. W., Gladwin, T. E., Hofmann, W., Salemink, E., and Ridderinkhof, K. R. (2013). Cognitive bias modification and cognitive control training in addiction and related psychopathology mechanisms, clinical perspectives, and ways forward. *Clin. Psychol. Sci.* 1, 192–212. doi: 10.1177/2167702612466547

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 February 2014; accepted: 29 July 2014; published online: 20 August 2014. Citation: Goel V (2014) Reason and less. Front. Psychol. ***5:****901. doi: 10.3389/fpsyg.2014.00901*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 01 July 2014
doi: 10.3389/fpsyg.2014.00625

Cognitive success: instrumental justifications of normative systems of reasoning

Gerhard Schurz*

Department of Philosophy, Heinrich-Heine University of Duesseldorf, Duesseldorf, Germany

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Jonathan St. B. T. Evans, University of Plymouth, UK
Igor Douven, University of Groningen, Netherlands

***Correspondence:**
Gerhard Schurz, Department of Philosophy, Director of Duesseldorf Center for Logic and Philosophy of Science, Heinrich-Heine University of Duesseldorf, Universitaetsstrasse 1, 40225 Duesseldorf, Germany
e-mail: schurz@phil.uni-duesseldorf.de

In the first part of the paper (sec. 1–4), I argue that Elqayam and Evan's (2011) distinction between normative and instrumental conceptions of cognitive rationality corresponds to deontological vs. teleological accounts in meta-ethics. I suggest that Elqayam and Evans' distinction be replaced by the distinction between a-priori intuition-based vs. a-posteriori success-based accounts of cognitive rationality. The value of cognitive success lies in its instrumental rationality for almost-all practical purposes. In the second part (sec. 5–7), I point out that the Elqayam and Evans's distinction between normative and instrumental rationality is coupled with a second distinction: between logically general vs. locally adaptive accounts of rationality. I argue that these are two *independent* distinctions that should be treated as independent dimensions. I also demonstrate that logically general systems of reasoning can be instrumentally justified. However, such systems can only be cognitively successful if they are paired with successful *inductive* reasoning, which is the area where the program of adaptive (ecological) rationality emerged, because there are no generally optimal inductive reasoning methods. I argue that the practical necessity of reasoning under *changing* environments constitutes a dilemma for ecological rationality, which I attempt to solve within a *dual* account of rationality.

Keywords: is and ought, normative accounts of rationality, means-end inference, cognitive success, general vs. locally adaptive rationality

INTRODUCTION: RECENT CRITICISMS OF NORMATIVE SYSTEMS OF REASONING IN PSYCHOLOGY

According to a common conception (Elqayam and Evans, 2011, p. 234), classical logic was the dominant normative standard of rational thinking in cognitive psychology until the 1960s. When psychologists discovered empirically that, in many domains, human reasoning did not accord with the principles of logic (e.g., Wason, 1966), these findings were interpreted as signs of human irrationality (cf. Evans, 2002). Beginning in the 1970s this interpretation came increasingly under attack by authors who demonstrated that deviations from classical logic can nevertheless be rational (e.g., Cohen, 1981; Gigerenzer, 1991; Oaksford and Chater, 1994). For example, when conditionals are uncertain, the optimal rules of conditional reasoning are no longer classical (see section Instrumental Justification of Deductive Reasoning). Some authors suggested that psychologists should adopt a different normative system, as an alternative to classical logic, such as, for example, Bayesian probability theory or decision theory (e.g., Oaksford and Chater, 1991, 2007). However, human reasoning has been observed to deviate from the norms of probability and decision theory, too (Kahneman and Tversky, 1972; Barbey and Sloman, 2007). Therefore other authors suggested that certain forms of "adaptive" or "instrumental" rationality do not presuppose any normative system at all; rather they can and should be studied in a purely descriptive way (e.g., Evans and Over, 1996; Gigerenzer et al., 1999). A clear exposition of this position is given in Elqayam and Evans (2011). I take this position as a starting point for my critical discussion of the notions of rationality that underlie the psychological debate on norms of reasoning.

According to Elqayam and Evans (2011, p. 234), *prescriptive normativism* is the view that human thinking *should* be evaluated against (the rules of) a normative system, *S*, and *ought* to conform to it, where S is a general system of reasoning such as logic, probability theory, or decision theory[1]. Elqayam and Evans launch three major criticisms against prescriptive normativism: (1) First, there are different mutually competing normative systems of reasoning, such as classical vs. non-classical logics, frequentistic vs. Bayesian probability theory, probability theory vs. fuzzy logic, etc. This leads to the problem of "arbitration," i.e., of deciding between different normative systems. For Elqayam and Evans it is more-or-less impossible to give an objective or unbiased approach to this problem, because normative systems understand their "norms" of reasoning as *fundamental* norms, being based on more-or-less *a-priori* intuitions which are not capable of further *rational* justification[2]. (2) Second, the endeavors of many psychologists to select one of these normative systems on *empirical* grounds are typically based on *is-ought fallacies*. According to a famous philosophical doctrine that goes back to Hume (1739/40, part 1, §1), it is logically impossible to infer an Ought from an Is. (3) Elqayam and Evans recognize instances of *ought-is fallacies* in psychological

[1] A forerunner is Evans and Over's notion of "rationality$_2$" (1996, p. 8).
[2] Elqayam and Evans (2011), p. 237f; in particular p. 277 in reply to Schurz (2011a).

research, in which psychologists infer incorrectly from their preference for a certain normative system S that a certain theoretical interpretation of people's empirical cognitive behavior is "correct," due to its coherence with the rules of reasoning prescribed by system S—a position which Elqayam and Evans call "empirical normativism" (Elqayam and Evans, 2011, p. 244, 234). In this way normative prescriptivism introduces *biases* which hinder empirical research.

Because of these problems Elqayam and Evans argue that psychologists of reasoning would be better off if they gave up normative prescriptivism and dispensed with appeals to any normative system whatsoever. They call this opposite position *descriptivism* and mention Gigerenzer and Todd's conception of *adaptive* (ecological) *rationality* as well as Evans and Over's *instrumental rationality* as prototypes of this position (Elqayam and Evans, 2011, P. 246f)[3]. For Evans and Over, a method of reasoning or decision-making is instrumentally rational if it is reliable and efficient *for* achieving one's *subjective goals* (1996, p. 8). Adaptive rationality is considered a kind of instrumental rationality which emphasizes the dependence of the optimal means of achieving one's goals on the given environment; so cognition can only be instrumentally rational if it is ecologically adapted. Like Evans and Over (1996), Todd and Gigerenzer (2012, p. 15) criticize the purported "a-priori" nature of normative systems and argue that the fitness of cognitive methods should be empirically tested in natural environments. Elqayam and Evans (2011, p. 247f) assure readers that their descriptivist position does not exclude normative recommendations entirely from the field of cognitive psychology. However, they argue that all that can be generally said about "rational thinking" (thereby quoting Baron, 2008) is that rational thinking is "whatever kind of thinking helps people to achieve their goals." This sounds very close to the pragmatist philosophy of William James.

In the following sections, I will try to embed the non-normativist positions of Evans and Over (1996), Elqayam and Evans (2011), and Todd and Gigerenzer (2012) into a more general philosophical framework. I will suggest replacing the conception of "normativist" vs. "instrumentalist" rationality by two independent distinctions: the distinction between intuition-based vs. success-based conceptions of rationality, and the distinction between logico-general vs. local-adaptive conceptions of rationality.

DEONTOLOGICAL vs. CONSEQUENTIALIST JUSTIFICATIONS OF NORMS IN THE CONTEXT OF THE IS-OUGHT PROBLEM

The distinction between normative vs. instrumental rationality is related to a standard distinction in meta-ethics: that between deontological vs. consequentialist justifications of norms. In *deontological* systems of ethics, the normative basis of justification consists of certain *fundamental norms*, which assert that certain general forms of action are *categorically* (i.e., unconditionally) obligatory or ethically *good in themselves*. In contrast, in *consequentialist* systems of ethics, actions are justified as normatively right because of the *value* of their consequences, or at least of those consequences that were foreseeable by the actor (Broad, 1930; Anscombe, 1958; Birnbacher, 2003; ch. 4). Consequentialist ethics are further divided into two groups: in *value-consequentialist* (or non-teleological) ethics, actions are normatively right because their consequences are ethically valuable, while in *teleological* ethics, actions are normatively right because their consequences promote the satisfaction of extra-ethical values, which consist in the *factual* goals of people, ultimately the avoidance of pain and achievement of pleasure (cf. Frankena, 1963, ch. 2). The most famous historical example of a deontological position is Kant's *categorial imperative*, which requires one to treat all morally relevant subjects equally and seems to be ethically right quite independently from its consequences. The most famous historical example of a teleological position is Benthem's and Mill's *utilitarianism*, for which an action is ethically right just in case it results in the "greatest happiness of the greatest number." While utilitarianism is an *altruistic* principle, its egoistic variant is *egoistic hedonism*, according to which an action is right for an agent if it maximizes the agent's *own* personal pleasure.

Let us discuss these ethical positions in the light of Hume's thesis that norms and ethical values cannot be logically derived from descriptive facts. Contemporary attempts to prove Hume's thesis by means of modern logic have faced surprising difficulties. These difficulties derive from the paradox of Prior (1960), which is based on two facts:

(i) From *purely descriptive* premises, for example ¬p (e.g., "I am not poor") one may derive *mixed* statements such as ¬p∨Pq, with "P" for "is permitted" (e.g., "I am not poor or stealing is permitted").
(ii) From the mixed statement ¬p∨Pq together with the descriptive premise p one can derive the *purely normative* statement "Pq."

Prior argued that if mixed statements count as descriptive, then (ii) counts as an is-ought inference, and if mixed statements count as normative, then (i) counts as an is-ought inference. So it seems that is-ought inferences result in either case (which constitutes Prior's paradox). The major insight that emerged from this paradox was that an adequate explication of Hume's is-ought thesis must be based on the *threefold* division of statements into purely descriptive, mixed, and purely normative. Based on this insight, Schurz (1997) proved that the following two versions of Hume's thesis hold in all standard logical systems of multi-modal first-order logic:

- (H1) No non-logically true purely normative conclusion can be derived from a consistent set of purely descriptive premises.
- (H2) Every mixed conclusion which follows logically from a set of purely descriptive premises is normatively irrelevant in the sense that all of its normative subformulas are replaceable by other arbitrary subformulas, salva validitate of the inference.

Thesis (H1) entails the *inverse* Hume thesis (H3) which says that no non-tautologous descriptive statement can be

[3]Cf. Gigerenzer et al. (1999), Todd and Gigerenzer (2012), and Evans and Over (1996).

logically inferred from a consistent set of purely normative premises[4].

In the light of the logical is-ought gap described by Hume, the explained positions of deontological, value-consequentialist, and teleological ethics describe the three major ways in which normative systems can be justified. Norms cannot be derived from facts alone, but they can be justified by deriving them from either (i) other norms having the status of *fundamental* norms (as in deontological ethics), (ii) fundamental ethical values, as in value-consequentialist accounts, or (iii) fundamental extra-moral values which are given by human goals or interests, as in teleological accounts. All three models of justification are based on a so-called *means-end* inference, which has the following form:

(1) *Means-end inference*:
Normative premise: A is a (fundamental) norm or value.
Descriptive premise: B is a necessary (or optimal) means for achieving A.
Normative conclusion: B is a derived norm or value.

This form of means-end inference is accepted as *analytically* valid within more-or-less all kinds of ethical theories, whether they are deontological, value-consequential, or teleological. Here "analytically valid" means "conceptually valid," i.e., "valid because of the meaning of the involved terms," but not "logically valid," i.e., "valid solely because of the meaning of the involved logical terms" (cf. Schurz, 2013, ch. 3.3–3.4). For example, "This is round, therefore it has no edges" is an analytically but not logically valid argument. Moreover, the analytical validity of the means-end inference holds only for necessary and optimal means to an end, but fails for sufficient means[5].

The second premise of the means-end inference, concerning the means-end relation, is a *factual* statement, expressing the results of empirical research. Thus, the means-end principle explains how the findings of empirical scientists can become practically relevant *without* committing an is-ought fallacy: empirical findings allow one to derive a multitude of derived norms from a small set of fundamental norms or values. The latter ones cannot be established by empirical science (following from Hume's is-ought thesis), but are given to the scientist by *extra-scientific* institutions, e.g., by politicians or by the society as a whole (Schurz, 2010, §6).

With help of means-end inferences one can only prove *hypothetical* (conditional) norms or values, i.e., *implications* of the form "if X is accepted as a norm or value, then Y is also normatively required or valuable"; but one can never justify *categorial* (unconditional) norms in this way (cf. Schurz, 1997, theorem 6, p. 132). For the latter purpose one needs additional premises. They come either in the form of *fundamental norms* or values, or, as in deontological or value-consequentialist accounts, in the form of factual interests of people together with *fundamental* ethical *is-ought* or is-value *bridge principles*, as in teleological theories. The fundamental bridge principle of hedonistic or utilitarian ethics, for example, states that "if the realization of a state of affairs p serves the interests of (some, most, or all) human beings, then p is valuable and ought to be realized." In deontological and value-consequentialist ethics, the most fundamental norms and values are assumed to be justified by *a-priori* intuition. However, teleological theories also contain such an element of a-priori intuition, in the form of a presupposed is-ought (or is-value) bridge principle. Bridge principles of this sort cannot be justified by logical inference (Hume's insight), nor by arguing that they are "valid by definition" (Moore's insight)[6]; they are often controversial and are accepted only in some but not in all ethical theories.

A-PRIORI INTUITION-BASED vs. A-POSTERIORI SUCCESS-BASED ACCOUNTS OF RATIONALITY

I will now try to connect the psychological distinction between normative and instrumental rationality to the philosophical framework of deontological, value-consequential, and teleological accounts in ethics, and evaluate the former distinction in the light of the latter. Obviously the position of Elqayam and Evans (2011) is a kind of teleological one, but exactly which one is not entirely clear, at least not for me. Nor is it prima facie clear which position is exactly criticized in their arguments against normative rationality—all non-teleological positions, or only certain ones? Let's see.

Elqayam and Evans understand the rules of a normative system of reasoning S as "evaluative" norms. They assume that evaluative norms are based on a-priori intuition, being unamenable to further justification. From a philosophical viewpoint, this view of (evaluative) norms is too narrow, since normative systems (be they deontological or value-consequentialist) *do* contain a multitude of *derived* norms, which are justified as (optimal or necessary) means to satisfying certain fundamental norms. For example, for Elqayam and Evans "poverty should not exist" is an evaluative norm (Elqayam and Evans, 2011, p. 236), but this norm is instrumental for the more fundamental norm that people should not suffer. Just the same is true for normative systems of reasoning: the fact that a general system of reasoning S such as logic or probability theory is accepted as a normative standard does not imply that reasoning in accord with S can only be justified by "a-priori intuition." Different ways of justifying the rules of logic or probability theory in terms of more fundamental norms, such as cognitive success or truth-conduciveness will be discussed in the section on General vs. Locally Adapted Rationality.

[4]Cf. Schurz (1997): for (H1) theorems 3–5 (p. 118, 121, 124), for (H2) theorems 1–2 (p. 92, 102), for (H3) prop. 7 (p. 74). For related work cf. Stuhlmann-Laeisz (1983), Pigden (1989), Galvan (1988); general introductions are Hudson (1969) and Pigden (2010).

[5]For example, for the purpose of letting fresh air into the room, tearing down a wall is a sufficient but neither a necessary nor an optimal means. Some people object that the means-end inference fails even for necessary (or optimal) means, since if the necessary means B for realizing the fundamental norm A is itself intrinsically bad, B should not be realized. However, in such a case it is unreasonable to accept A as a fundamental norm. Thus, although this objection points to an important constraint on fundamental norms, it fails as an argument against the means-end inference.

[6]Cf. Moore's famous "open question" argument against the "naturalistic fallacy" of defining "Ought" by "Is" (1903, p. 15f).

There is, however, a philosophical position to which Elqayam and Evan's criticism does indeed apply. A well-known example of this position is Cohen's account of rationality (1981). For Cohen, rules of logical reasoning such as Modus Ponens or Modus Tollens are based on a-priori intuitions about correct reasoning. If human reasoning deviates from the rules of logic, this could mean for Cohen that these people have different a-priori intuitions about correct reasoning. So they are not irrational, but their reasoning is merely based on a different norm of rationality. Cohen understands his position as a generalization of Goodman's and Rawls' coherentistic conception of a "reflective equilibrium," which involves the balancing of general intuitions about correct rules and particular intuitions about rule-instances (Goodman, 1955; Rawls, 1971). I propose to call this family of positions "a-priori intuition-based" conceptions of rationality, which I set in opposition to *a-posteriori success-based* conceptions of rationality.

Intuition-based conceptions base rationality on a motley "stew" of intuitions, including intuitions about the correctness of cognitive rules such as Modus Ponens in logic or Bayes' theorem in probability theory. These intuitions are "subjectively a-priori" in the sense that they are taken as primitively given, incapable of further justification, although they can vary between different subjects. For example, religious people would consider different rules of reasoning as "intuitively rational," compared to non-religious people. It is therefore unavoidable that this conception of rationality must lead to a strong form of *cognitive relativism*, which has in particular been worked out by Stich (1990).

The notion of "prescriptive normativism" as characterized by Elqayam and Evans (2011) or Evans and Over (1996, p. 8) seems to correspond to a-priori intuition-based conceptions rationality. I agree with Elqayam and Evans' criticism of these positions: they take unreliable subjective intuitions as sacrosanct and thereby hinder rational criticism and scientific progress. However, the opposite of a-priori intuition-based conceptions of rationality are not "descriptive" conceptions of rationality (whatever these may be), but rather a-posteriori success-based conceptions of rationality, which evaluate systems of reasoning in terms of the cognitive *value* of their *consequences* in the given *environment*.

The emphasis of the *local adaptivity* of successful reasoning systems, i.e., the dependence of their value on the environment in which they are applied, is a central insight of the research program of ecological rationality. Todd and Gigerenzer (2012, p. 15) write: "We use the term *logical rationality* for theories that evaluate behavior against the laws of logic or probability rather than success in the world," while "The study of ecological rationality is about finding out which pairs of mental and environment structures go together." Todd and Gigerenzer's understanding of "logical rationality" matches our notion of a-priori intuition-based accounts of rationality, and their notion of ecological rationality fits with our understanding of a-posteriori success-based accounts, except that I support a dualist standpoint (similar to Evans, 2003), according to which not *only* locally adapted but also certain general reasoning methods can be justified in this success-based way (see the section on General vs. Locally Adapted Rationality).

In the light of contemporary epistemology (cf. Greco and Turri, 2013), a-priori intuition-based accounts are *internalist* accounts of rationality, because they understand the rationality of a cognitive act as an internal property of the underlying cognitive process of the agent, independent from the environment. What these accounts have in common with deontological ethics is that they evaluate the moral rightness of an act solely based on the properties and intentions of the actor at the time of acting, independent from its consequences. In contrast, a-posteriori success-based accounts are *externalist* accounts of rationality, inasmuch as the success of a cognitive act depends on its consequences in the given environment; this is what these rationality accounts have in common with *consequentialist* ethics.

I do not deny that a-posteriori success-based accounts of rationality also involve *some* elements of intuition. But their intuitive elements can be narrowed down to a few fundamental intuitions about human *goals* whose realization are assumed to be *valuable* (which is a fact-value bridge principle of the explained sort). What a-posteriori accounts reject is reliance on *epistemic correctness intuitions*, i.e., intuitions about the epistemic correctness or plausibility of rules of reasoning. In a-posteriori accounts, all epistemic correctness claims of this sort have to be justified by means-end inferences, which attempt to show that the respective rules are instrumental for attaining the assumed goals in the assumed class of environments.

The notion of *success* contains an objective component (success *in* the given environment) as well as a subjective component (success *for* assumed goals). As long as the "goals" *for* an action are not specified, it is prima facie *unclear* what is meant by "success." Todd and Gigerenzer avoid making any general statement about what the success of cognitive methods consists in. In all of their experiments, however, they assume that the success of a cognitive method increases with the frequency of its "correct" or empirically true inferences (or predictions), and decreases with the cognitive costs of the method, in terms of necessary information search and computation time. This understanding of "cognitive success" is widely accepted in cognitive science. Philosophers often neglect the dimension of cognitive costs and define *truth-conduciveness* (attainment of true and avoidance of false beliefs) as the fundamental epistemic goal (David, 2005). A prominent variant of this position is *reliabilism* (cf. Goldman, 1986; Schurz and Werning, 2009)[7].

TRUTH-CONDUCIVENESS AND COGNITIVE SUCCESS: INSTRUMENTALLY RATIONAL FOR ALMOST ALL PURPOSES

I suggest that the fundamental *goal* of cognitive methods in a-posteriori accounts of rationality should be characterized as the maximization of cognitive success, in the explained sense of finding many possibly relevant truths with little cognitive effort. While this position would find many friends within contemporary epistemology, the notion of "truth" seems to be less popular in cognitive psychology. In their reply to Schurz (2011a), Elqayam and Evans (2011, p. 278f) reject truth as the general goal of reasoning, in favor of an unspecific notion of "instrumental rationality," which is relativized to arbitrary goals. Before we discuss this position, let us analyze the goal of truth-conduciveness

[7]Goldman (1986) is an exception among epistemologists inasmuch as he also discusses cognitive costs.

or cognitive success in the light of the preceding discussion of positions in ethics. Two general views are possible: (1) One may understand truth-conduciveness as a fundamental epistemic value, incapable of further justification; this understanding corresponds to a value-consequentialist position. (2) One may deny that truth-conduciveness is an "intrinsic value," but understand the value of cognitive success instrumentally, in terms of its usefulness for the achievement of some given extra-epistemic (or practical) purposes, whatever these purposes may be. The latter viewpoint corresponds to the *teleological* position of *instrumental* rationality, which underlies the views of Elqayam and Evans, Gigerenzer and Todd, and perhaps the majority of psychologists.

First of all, I wish to point out that although instrumental norms or values are *hypothetical* in the sense explained in sec. 2, they are not "descriptive," but nevertheless possess normative or evaluative content. Recall that, although the second premise of the means-end inference (1) is descriptive, its conclusion is normative or evaluative: it inherits this status from the first premise which asserts that something is a fundamental norm, value, or goal. This is also true when the fundamental value is given by the factual subjective goal of one or many persons (together with a fact-value bridge principle[8]). For example, if it is a fundamental value *for me* to protect the environment, then it is a derived value for me to support Greenpeace.

Secondly, Elqayam and Evans' conception of instrumental rationality may be relativized to *any purpose whatsoever*. They endorse the view that "rational thinking is whatever kind of thinking best helps people achieve their goals" (Elqayam and Evans, 2011, p. 248). Let us ask: doesn't this position imply that rationality, itself, is entirely relative? On closer inspection, the notion of instrumental rationality is semantically ambiguous. At least three different conceptions of instrumental rationality exist in the literature: instrumental rationality as (i) technocratic rationality, (ii) goal-relative rationality, or (iii) general all-purpose rationality. While (i) maintains that instrumental rationality is ideologically biased (Habermas, 1966), (ii) maintains that it is entirely relative: there are as many kinds of instrumental rationality as there are different kinds of human goals (Stich, 1990). Only position (iii)—which I attribute to Evans and Over (1996, p. 8)—maintains that instrumental rationality is general and non-relative.

Is it an unavoidable consequence of the notion of instrumental rationality that it is goal-relative? Do we have for each kind of goal a separate account of rationality? Do environmentalists, warriors, and taxi drivers, etc. each employ different methods of rational reasoning? *This seems to be entirely wrong.* In this section, I present a simple argument that shows that there is a form of rationality that is instrumental for almost all purposes: this form of rationality is contained in the idea of truth-conduciveness in the explained sense. *This* is the practical reason why it makes sense to *separate* epistemic from non-epistemic goals, and regard the satisfaction of epistemic goals as a good, independent of the practical goals which one actually pursues. I say "for almost all purposes" because there are some important exceptions which I will discuss later. First let me briefly explain—or since this is so obvious, I should better say: recall—why truth-conduciveness is all-purpose instrumental.

Maximizing the utility of one's practical actions (whatever they are) is usually explicated in terms of a *decision situation*. The task is to choose that action among a possible set of competing actions which has the maximum expected utility. Therefore, each decision problem can be reduced to a prediction problem whose task it is to predict which of the possible actions will lead to a maximal expected payoff (in Schurz, 2012 this method is used to reduce action games to prediction games). To predict the expected payoff of the available actions, it is necessary to predict the environmental conditions under which the actions will take place, and the consequences of each action under these conditions. In this way, practical success in a given decision problem depends on cognitive success in a corresponding prediction problem. Therefore, increased success in one's predictions, as measured by the goal of truth, will by and large lead to increased success in one's practical actions, *independently of the goals which one pursues*.

Elqayam and Evans reject truth-conduciveness as the supreme cognitive goal for reasons which do not really conflict with my arguments. They understand the notion of truth in a much more "metaphysical" and less practical and empirical sense than I do. For example, they argue that "cognitive representations are viewed not as veridical, but as fit for purpose" (Elqayam and Evans, 2011, p. 278). This is nothing but the teleological position explained above. They continue with remarking that there is no "true picture of the world which our eyes and brains deliver faithfully to us. There is a mass of information in light, which could be interpreted and constructed in many ways. In addition, our visual systems have clear limitations." From a philosophical standpoint all of this is obviously true. But this only means that we never know the *complete* truth ("true picture") of our environment, and that our cognitive models are never free from *simplification* and *error*. However, all that counts for practical success is true information about practically relevant questions, for example whether or not it will rain tomorrow, or whether the value of a given share will go up or down. By "cognitive success" I do *not* mean the achievement of fancy metaphysical truths, but (at least primarily) the achievement of *empirical* (i.e., possibly observable) truths, which are of possible relevance for our practical success. This position is not far from Elqayam and Evans, who infer from their considerations that "cognitive representations are only veridical to the extent and in the manner required to serve our goals." In conclusion, I am inclined to think that Elqayam and Evans' "instrumental rationality" along with Todd and Gigerenzer's "adaptive rationality" and Evans and Over's "rationality$_1$" can be subsumed under the family of a-posteriori conceptions of rationality, which evaluate rationality in terms of their cognitive success in the sense just explained.

Let me finally mention the *big exception* to the all-purpose instrumentality of truthful beliefs. Our beliefs may have certain *direct* effects on us that are quite independent from their truth value. If I believe that a beloved person will visit me in an hour, then this belief makes me happy for the next hour, quite independently of whether or not this person actually comes. Schurz (2001a) calls these effects the *generalized placebo effects* of our

[8]This bridge principle says: "If person X has goal A, and A is not in conflict with other goals of X, then A's realization is valuable for X" (cf. Schurz, 1997, sec. 11.7).

beliefs. Placebo effects have been extensively studied in the area of medicine and pharmaceutics. For example, the mere belief in the effectiveness of a sleeping pill accounts for more than 50% of the success of a real sleeping pill. More generally, positive illusions have positive effects on a person's physical and mental health (Taylor, 1989, p. 49, 88ff, 117ff). Particularly effective in this respect are *religious* beliefs. Because of their selective advantages, generalized placebo effects are to some certain extent built into our cognitive processes and are the reason for certain cognitive "biases" that have been discovered in the heuristics-and-biases research in psychology. Piatelli-Palmarini (1996) classifies these cognitive biases in seven groups, where at least three of them are the result of genetically selected placebo effects: overconfidence, hindsight bias, and self-righteous bias.

Placebo effects are *real* and *useful* effects, being produced by one's strong belief in some usually false state of affairs, for example in one's own superiority or in the existence of a safeguarding God. However, placebo effects break down as soon as one comes to believe the truth: the resistance of my body against cancer decreases when my doctor tells me that my chances to survive are small (etc.). In conclusion, placebo effects are the big exception to the all-purpose instrumentality of true beliefs.

Let me emphasize that my comment concerning placebo effects is only intended to direct attention to this problem, but not to offer an adequate treatment (or "solution") as the latter project would exceed the scope of this paper. Rather than offer a solution, I want to conclude my discussion of this problem with the following remark. The unjustified faith in one's beliefs upon which the placebo effect rests is at the same time practically dangerous: it often leads to a dogmatic belief system which resists revision through the scientific procedures of critical testing, and promotes tendencies to solve conflicts by fiat or violence instead of rational reflection. Despite the beneficial aspects of placebo effects, eliminating vulnerability to placebo effects is a *price* that must paid as a means to acquiring a scientific as opposed to a magical belief system—a price that is worth paying, given the general value of truth beliefs for practical action and the dangers of dogmatic belief.

GENERAL vs. LOCALLY ADAPTED RATIONALITY: NOT A NORMATIVE BUT A DESCRIPTIVE QUESTION

In the preceding sections, I investigated the normative side of rationality. I distinguished two accounts of rationality, a-priori intuition-based vs. a-posteriori success-based. To some extent this distinction reflects Elqayam and Evan's (2011) distinction between prescriptive normativism and descriptive instrumentalism, and Todd and Gigerenzer's (2012) distinction between logical and ecological rationality. Both accounts contain *some* normative elements (in this respect I disagree with Elqayam and Evans), which are, of course, much stronger within intuition-based than within success-based accounts. In the former accounts, the normative elements derive from a mixed bag of a-priori intuitions, while in the latter accounts the only element of intuition concerns the acceptance of cognitive success, i.e., practically relevant truthfulness, as the fundamental cognitive goal, which is justified by its almost-all-purpose instrumentality for practical success.

There is, however, a second distinction, that between *logically general* vs. *locally adaptive* accounts of rationality. Todd and Gigerenzer (2012, p. 15) as well as Elqayam and Evans (2011) equate this distinction with the one between a-priori and a-posteriori accounts: for them logically-general accounts would be normatively justified in an a-priori manner, while locally adapted accounts are a-posteriori justified by their cognitive success in a given kind of environment. In my view, these two distinctions should be treated as two *independent dimensions* of classification, for the following reasons. *Firstly*, logico-general systems of reasoning can also be instrumentally justified, by their a-posteriori success in regard to—not specific but varying—environments and cognitive tasks. *Secondly*, the locally adaptive view of rationality can also be quite explicitly normative: Todd and Gigerenzer (2012) is full of recommendations to use frugal locally adapted heuristics instead of general logical tools. *Thirdly*, a local and special-purpose-related cognitive method, such as Kahnemann and Tversky's availability heuristics, can also be justified by a-priori intuitions: this is the way that such heuristics are justified within Cohen's (1981) "reflective equilibrium" account of cognitive rationality.

While the question of deciding between a-priori intuition-based vs. a-posteriori success-based accounts is a *meta-normative* question, concerning the way normative recommendations can be justified, the question of whether logico-general or locally adapted reasoning methods are more successful is a *descriptive* question, that only can be decided by computational and empirical means. To avoid misunderstanding: this question cannot be decided by finding out which cognitive methods are implicitly used by ordinary people when they reason. This would involve an is-ought fallacy, since we should not expect the actual reasoning of humans to always be cognitively successful or optimal. However, the cognitive success of reasoning methods can be studied by means of logical arguments, by mathematical theorems, and by empirical investigations of their performance in simulated and real-world environments. In the following subsections, I will sketch some typical success-based justifications of cognitive methods, both of logico-general methods that are prominent in philosophy, and of locally adapted reasoning methods that have been promoted by defenders of ecological rationality. I will show, in each case, that a closer inspection of these success-based justifications reveals that the cognitive success of the respective methods is limited to certain situations. The presented justifications of the "competing" cognitive methods do not really contradict each other; only their uncritical generalization as "autocratic paradigms" leads to mutual conflict.

INSTRUMENTAL JUSTIFICATION OF DEDUCTIVE REASONING

The standard justification of (classical) deductive reasoning consists in the provable fact that this kind of reasoning *preserves truth with certainty*: in all possible worlds in which the premises of a deductive argument are true, the conclusion of the argument is also true.

First of all, let me try to remove a misunderstanding which is apparently involved in some arguments that set logical and probabilistic accounts of reasoning in opposition to one another. For example, Elqayam and Evans (2011, p. 278) infer from the

fact that Bayesian updating is only possible if the probabilities are non-extreme (different from 1 and 0) that truth-preserving deductive inference is in conflict with Bayesian belief-updating. In this argument they equate the truth of a premise with its having an epistemic probability of 1. However, these two things are entirely different: (1) Obviously, truth is different from having an epistemic probability 1, since something can be true despite the fact that I don't believe it, and vice versa. (2) Further, believing that something is true is not the same as believing it with probability 1, because if I am a fallibilist then I will believe that the proposition I believe could be false, which means that I assign to them a high but not maximal probability. Moreover, knowing that a set of premises entails a certain conclusion can be cognitively useful even if my degree of belief in these premises is non-maximal. It is a straightforward theorem of probability theory that the probability of the conclusion of a valid deductive inference must be at least as high as the probability of the *conjunction* of its premises. Many further theorems of this sort have been proved in the literature, for example, that the uncertainty (i.e., 1 minus the probability) of the conclusion must always be greater than or equal to the sum of the uncertainties of the premises (Suppes, 1966, p. 54). In conclusion, the account of deductive reasoning is *not at all in conflict* with the account of probabilistic reasoning (see also the section Instrumental Justification of Probabilistic Reasoning on this point).

But let us ask: what does the truth-preserving nature of deductive reasoning imply regarding the cognitive success and usefulness of deductive inferences? In order to make cognitive use of a deductive inference, two conditions must be satisfied:

(2) *Conditions for the cognitive usefulness of a deductive inference:* (a) It must be possible for a person with "normal" cognitive abilities to achieve reliable beliefs about the truth of each premise, without (b) that the achievement of this belief relies itself on the person's belief in the truth of the conclusion.

Only if these two conditions are satisfied, can the cognitive process of drawing the deductive inference produce a new belief for the given person, which then is at least as reliable (i.e., probable given the evidence) as the conjunction of its premises. Condition (2a) entails that the premises must be consistent. Condition (2b) is violated, for example, in trivial logical inferences such as "p and q, therefore p," since all persons with normal cognitive abilities will, in the moment in which they start to believe a conjunction of two beliefs, believe each of its conjuncts. This is not the case in more complicated cases of deductive inference: for example, no cognitively normal person will believe that there are infinitely many prime numbers, from the moment that she begins to understand and believe the axioms of Peano arithmetic. In cases of this sort, deductive proofs produce new cognitive insights and, hence, are cognitively useful.

In the last example, belief in the premises (Peano's axioms of arithmetic) are believed based on mathematical "intuition" or postulate. In empirical applications, knowledge of the premises must be based on empirical evidence. Here we meet a further condition for cognitively usefulness. Nontrivial cognitive inferences usually involve conditionals (implications), which in classical logic are *material* conditionals "p→q," whose truth-table coincides with "¬p or q." As a consequence, "p→q" follows deductively from "¬p" and from "q." I call a material conditional *trivially verified* if the belief in it is justified either by the belief in the negation of its antecedent (¬p) or by the belief into its consequent (q). One can easily see that deductive inference from trivially verified conditionals cannot be cognitively useful:

(3)		*Trivial verification of 1st premise by*:
(a)	Modus p→q	¬p: Then verification of 2nd premise is impossible
	Ponens: p	q: Then conclusion is already known: inference trivial
	q	
(b)	Modus p→q	q: Then verification of 2nd premise is impossible
	Tollens: ¬q	¬p: Then conclusion is already known: inference trivial
	¬p	

Similar considerations apply to more complicated inferences (see footnote 10), for example to inferences from disjunctions, such as disjunctive syllogism: "p∨q, ¬p, therefore q." It follows that deductive inferences can only be cognitively useful if their conditional or disjunctive premises are not known by trivial verification. In empirical (non-mathematical) domains the standard way of justifying a *singular* material conditional without knowing the truth value of its if-part and then-part is to infer it from a corresponding *general* conditional, which is in turn *inductively* inferred from the empirical evidence. For example, when I believe that "if Jonny promises to come, then he will come," I don't believe this because Jonny didn't promise to come or because he actually came, but because I inferred this prediction from his promise-keeping behavior in the past.

In classical logic one can only express *strictly* general conditionals, which don't admit of exceptions and have the form "For all x: Fx→Gx" (with "Fx/Gx" for "x has property F/G"). With their help, the inference in (3a) is transformed into the following:

(4) *Modus Ponens from the instantiation of a strictly universal conditional*:

For all x: Fx→Gx	Nontrivial confirmation of the 1st premise by a sample of Fs all of
a is an F	which are Gs, where this sample doesn't contain individual a.
a is a G	

It follows from these considerations that in application to empirical knowledge, the successful cognitive use of deductive inferences is usually[9] restricted to *situations* which satisfy the following two conditions:

[9]A generalization to deductive inferences of arbitrary kind is possible by the following consideration: Statements which are verifiable by observation have the form of closed literals, i.e., unnegated or negated statements of the form (¬)Fa, or (¬)Rab, etc. (where F, R, etc. are primitive non-logical predicates).

(A) The inference contains at least one conditional (or disjunctive) premise which is explicitly or implicitly[10] general and can only be confirmed by an inductive inference, and
(B) the generality of this premise is strict (i.e., exceptionless).

Condition (A) implies that without the simultaneous capacity to reason inductively, deductive inferences are of almost no use in empirical domains. So condition (A) alone is sufficient to refute the view that deductive logic is an all "all-purpose" system of reasoning: although its inferences are truth-preserving in all possible situations (or worlds), they are not cognitively useful in all possible situations[11], but only in those situations in which deductive reasoning competence is *paired* with success in inductive reasoning.

I assume that for the majority of readers nothing of what I have said is substantially new. But if this is so, I cannot understand how one can seriously regard deductive logic as *the only* normative standard of reasoning. I guess that many of the hegemony claims made on the behalf of given "normative systems" are more the result of power struggles between Kuhnian "paradigms" than of rational reflection.

Condition (B) is an equally severe restriction on the cognitive use of deductive logic. Apart from laws in classical physics, there are not many true and strictly general laws in the empirical sciences. Most empirical conditionals are *uncertain* and admit of exceptions; they have the form "Most Fs are Gs," or "Normally, Fs are Gs" (cf. Schurz, 2001b, 2002). Conditionals of this sort are usually reconstructed as expressing high conditional probabilities. Reasoning with them requires probabilistic systems, either in the form of a conditional logic based on a probabilistic semantics (cf. Adams, 1975; Schurz, 2005; Schurz and Thorn, 2012; Thorn and Schurz, 2014), or within the full system of mathematical probability theory (see the section Instrumental Justification of Probabilistic Reasoning). Experimental investigations of reasoning have confirmed that people frequently understand uncertain conditionals in the sense of high conditional probabilities (Evans et al., 2003; Schurz, 2007). We will see below, however, that the application of probability theory (or of more advanced mathematical theories) to empirical domains is also only cognitively successful if it is paired with inductive reasoning mechanisms which can provide empirical confirmation for the general premises.

Note that conditions (A) and (B), above, are less restrictive than it may seem. First of all, conditions (A) and (B) do not apply to mathematical domains, where strictly general premises are given by axiomatic stipulation. No wonder, therefore, that deductive inferences are most intensively used in the mathematical sciences. Secondly, the fact that inferences from uncertain scientific laws require the use of probability theory does *not* make deductive logic disappear, because probability theory is usually formalized within standard type-free (Zermelo Fraenkel) set theory, which contains in its core the full power of deductive logic, which is needed, for example, to prove probability theorems from probability axioms. Replacing mere logic by advanced mathematics only breaks the autocracy of logic, but not its omnipresence: all higher-level mathematical theories still contain logic in their core. In conclusion, standing on its own legs deductive logic is highly useful in mathematical domains. In empirical applications, however, its cognitive use is confined to situations in which deductive reasoning is combined with the results of inductive reasoning procedures.

INSTRUMENTAL JUSTIFICATION OF PROBABILISTIC REASONING

The instrumental justification of probabilistic reasoning in terms of cognitive success depends on the assumed conception of probability: *statistical* (objective) or *epistemic* (Bayesian, subjective)[12]. The *statistical* probability of a property or event-*type* Fx, p(Fx), is the *limit of its relative frequency* in an underlying *random sequence* consisting of the consecutive outcomes of a random experiment (important founders are Von Mises, 1964, and Fisher, 1956). On the other hand, the *epistemic* probability of a particular state of affairs or event-*token* Fa, P(Fa), is the degree of belief, to which a given rational subject, or all subjects of a certain rationality type, believe in the occurrence of the event (important founders are Bayes, 1763; Ramsey, 1926; De Finetti, 1937).

The standard justification of Bayesian (i.e., epistemic) probabilities is their interpretation as fair betting quotients. Ramsey and de Finetti proved that a bettor's fair betting quotients satisfy the (standard Kolmogorovian) probability axioms if and only if they are *coherent* in the sense that there is no finite class of fair bets which under all possible circumstances lead to a total loss for the bettor. According to this view, the cognitive usefulness of coherent degrees of beliefs consists in the avoidance of sure loss, independent from the given environment. Although I do not deny that this form of probabilistic consistency is of "some use," it is certainly not enough for truthful prediction or successful action in the *actual* world. The definition of coherent fair betting quotients refers solely to the *subjective mental state* of the betting persons, but it need not reflect the true frequencies of the bet-on events. Take for example a subjective Bayesian who offers odds of 1:1 that she will roll a six with a normal die, and considers the bet fair, i.e., she is willing to accept the opposite bet at 1:1 that she won't roll a six. The Bayesian remains coherent even after she has lost her entire fortune. She may be puzzled that while everybody has readily accepted her bet, nobody has accepted the counterbet, but she can't explain why she of all people has lost everything while others have made their fortune, *as long as* she

However, deductive inferences among literals necessarily fail to meet condition (2b) of cognitive usefulness, because a literal follows from a set Δ of literals if and only if it is an element of Δ [cf. (Schurz, 2011b), sec. 5.1, (5)].

[10]The conditional premise $Fa \rightarrow Ga$ is said to be *implicitly general* if it is justified by an argument which justifies a corresponding conditional $Fa_i \rightarrow Ga_i$ *for every other individual constant* a_i. For example, $Fa \rightarrow Ga$ is implicitly general if it is deductively inferred from the explicitly general premise $\forall x(Fx \rightarrow Gx)$, or if it is inductively inferred from sample information of the form $\{Fb_1 \rightarrow Gb_1, \ldots, Fb_n \rightarrow Gb_n\}$.

[11]My notion of a *situation* includes both (a) an objective (subject-independent) environment, and (b) a constellation of subjective facts concerning the given cognitive task and cognitive resources.

[12]We confine our discussion to these two most important conceptions of probability. Further probability concepts which we cannot discuss here are objective single case probabilities and logical probabilities. Cf. Gillies (2000, ch. 3.13).

doesn't consider the frequentistic chances of the type of event she has been betting on. This shows that Bayesian coherence provides at best a *minimal* condition for rational degrees of belief, which is, however, too weak to exclude irrational betting behavior from an objective point of view.

In other words, subjective degrees of belief can *only* be cognitively successful if they are related to statistical probabilities. The most important connection between subjective and statistical probabilities is expressed by a principle that goes back to Reichenbach (1949) (cf. Schurz, 2013, p. 132):

(5) *Principle of narrowest reference class:* the subjective probability P(Fa) of a single event Fa is determined as the (estimated) *conditional* statistical probability p(Fx|Rx) of the corresponding type of event Fx in the *narrowest* (nomological) reference class Rx, within which we know a lies (i.e., that Ra is true).

The principle of the narrowest class of reference (also called the "statistical principal principle"; Schurz, 2013, p. 262) is widely used both in everyday life and in the sciences. If we want to determine the subjective probability that a certain person will take a certain career path (Fa), then we rely on the characteristics of this person which are known to us as the narrowest reference class (Ra), and on the statistical probability that a person x with the characteristics Rx will take this career path (p(Fx|Rx)). The weather forecast "the probability that it will rain *tomorrow* is 3/4" has, according to Reichenbach's principle, the following interpretation: the statistical probability that it will rain on a day which is preceded by similar weather patterns as that preceding today is 3/4. Unterhuber and Schurz (2013, sec. 4.3) argue that Oaksford and Chater (2007), too, seem to accept a principle of this sort.

Bayesian probabilities can only be truth-conducive and cognitively useful if they are connected with statistical probabilities. Only if we can reliably predict the true success probabilities of our actions can our actions be useful. However, all knowledge about statistical probabilities must be inferred from observations of past instances or samples by means of *inductive* inferences. So our conclusion concerning the cognitive usefulness of probability theory is similar to our conclusion for deductive logic: in application to empirical domains, probabilistic reasoning is only useful if it is combined with the capacity for successful inductive inference. In fact, there exists a manifold of accounts which explicate different forms of inductive inference in probabilistic ways—for example, Fisher's, and Neyman and Pearson's account of statistical tests, Fisher's account of statistical inference based on confidence intervals, the approach of Bayesian statistics based on the updating of prior distributions, etc. (for an overview cf. Schurz, 2013, ch. 4). Although it is not possible to enter into the details here, we note that all of these accounts assume *special* principles or rules that go beyond the basic axioms for coherent probabilities and correspond to different forms of inductive inference.

INSTRUMENTAL JUSTIFICATION OF INDUCTIVE INFERENCE: LOCALLY ADAPTED METHODS

In the two preceding subsections we have seen that classical logic and probability theory are far from being "all-purpose" cognitive tools. To be sure, deductive inferences are truth-preserving and coherent probabilities avoid sure-loss, but beyond that, the two reasoning systems can only be successful in empirical applications if they are paired with successful inductive inferences. It is the very domain of inductive inferences, however, in which no universally reliable method, nor even a universally optimal method, exists. Negative results of this sort basically go back to the insights of the philosopher David Hume and have more recently been proved in the areas of formal learning theory (Kelly, 1996), machine learning (Cesa-Bianchi and Lugosi, 2006), and meta-induction (Schurz, 2008; Vickers, 2010, §6.3). This is not to deny that in the area of inductive prediction there are a variety of positive results, but they either hold only under restrictive conditions, or they hold only in the "infinitely long run," and tell us nothing about the cognitive success of a respective method in practically relevant time. So very naturally, inductive prediction tasks have been the domain in which the paradigm of *locally adaptive* or *ecological* rationality has emerged, which has been developed, among others, by Gigerenzer, Todd, and the *ABC research group*[13]. These researchers show, based on comparative investigations of the success of different prediction methods, that simple prediction heuristics are frequently more successful than more general and computationally costly prediction mechanisms, following the slogan "less can be more." Gigerenzer et al. (1999) have studied several different heuristics at different levels of generality. In this subsection I focus my discussion on the performance of one of these prediction rules, known as "take-the-best" (TTB).

The prediction tasks studied within the ABC research group have the following format: prediction methods are based on so-called *cues* C1,. . .,Cn, which are themselves *predictive indicator*s of a criterion variable X whose values or value-relations have to be predicted. Each cue has a given *probability* of predicting correctly, conditional on its delivering any prediction at all. This conditional probability is called the cue's ecological *validity*. In one of the typical experiments, the task was to predict which of two German cities has a higher population, based on binary cues such as (C1) is it a national or state capital?, (C2) does it have a first division soccer team?, etc. In experiments of this sort, a cue (Ci) delivers a prediction if it "discriminates" between the two compared objects: if the cue difference is $+1$ (value 1 for city A and 0 for city B), the cue predicts $X_A > X_B$ (city A is larger than city B); if the cue difference is -1 (value 0 for city A and 1 for B), it predicts $X_B > X_A$ (city B is larger than A), and otherwise it fails to predict.

For each item (i.e., pair of cities), the strategy TTB predicts what the cue with the highest ecological validity predicts, among all cues which deliver a prediction for the given item. The frugality of this strategy consists in the fact that for each item it bases its prediction on only one cue (that with the highest validity). In contrast, more complex strategies predict a certain (mathematical) combination of the predictions of all cues. For example, the strategy called "Franklin's rule" predicts

[13] Cf. Gigerenzer et al. (1999) and Todd and Gigerenzer (2012). "ABC" stands short for the "Center for Adaptive Behavior and Cognition" at the MPI for Human Development in Berlin.

according to a weighted average of the cue differences of all discriminating cues, where the weights are determined by the (normalized) validities of the discriminating cues (Gigerenzer et al., 1999, part III). If this weighted average is greater (or smaller) than 0.5, Franklin's rule predicts $X_A > X_B$ (or $X_A < X_B$, respectively). A still more complex prediction method is linear (or logistic) regression: this method predicts a linear (or logistic) combination of the cue differences with optimal weights which minimize the sum of squared distances between the actual value of the item (which takes $+1$ if $X_A > X_B$ and -1 if $X_A < X_B$) and the predicted linear (or logistic) combination of cues (Gigerenzer et al., 1999; Rieskamp and Dieckmann, 2012).

It can be proved that both regression methods have equally maximal predictive success among all linear combinations, if their weights are fitted to 100% of all items of the underlying population (in our example all pairs of cities). In practice, however, the weights are estimated from so-called *training sets*, which consist of random samples of varying size (e.g., 20% of all items). Likewise, Franklin's rule and TTB estimate the validities of the cues from their validities in training sets. This is the point where the advantage of frugal strategies such as TTB comes in. Regression methods, and to some extent also Franklin's rule, suffer frequently from the problem of *overfitting*: they fit the weights or validities to random accidentalities of the sample which disappear 'in the long run,' when the samples size approaches the population size (cf. Brighton and Gigerenzer, 2012, Figures 2–1; Rieskamp and Dieckmann, 2012, p. 198f, Figures 8–1, 8–2). Based on simulated and real data, Rieskamp and Dieckmann (2012) arrive at the result that linear weighting methods tend to be better than TTB in environments of low redundancy, with little (unconditional) correlations between the cues' predictions, while in high redundancy environments TTB tends to be better than weighting methods (for small learning samples) or equally good (for large learning samples)[14]. However, one can show that Rieskamp and Dieckmann's generalizations from their empirical results are not always correct. Schurz and Thorn (in review) construct environments in which weighting rules are superior in spite of redundant cues, as well as environments in which TTB is superior in spite of non-redundant cues (see Appendix). In the next section we will see that there is a systematic reason for the difficulty of providing simple rules that characterize the class of environments in which frugal prediction methods such as TTB beat complex prediction methods such as Franklin's rule or regression.

A DILEMMA FOR ECOLOGICAL RATIONALITY, OR WHY A DUAL ACCOUNT IS NEEDED

The success of any locally adapted prediction method depends on its being applied in the "right" *environment*. However, biological organisms, and especially humans, frequently face *changing* environments. Within such environments, one needs strategies that *select* for each relevant environment a method, or a *combination* of methods, that performs as well as possible in that environment. Following Rieskamp and Otto (2006, p. 207), I call this the *strategy selection* problem.

Researchers within the adaptive rationality program acknowledge the importance of the strategy selection problem. For Todd and Gigerenzer (2012 p. 15), the study of ecological rationality centers around the question of which heuristics are successful in which kinds of environments. They propose a list of simple rules which indicate, for each of their studied heuristics in which kind of environment it may be successfully applied, and in which it may not (Todd and Gigerenzer, 2012, Table 1.1). On closer inspection, however, their rules are problematic, either because their application requires information that is unlikely to be available, or because the rules are not always correct. For example, the recognition heuristic ("base your prediction on that cue which is best recognized by you") is said to be ecologically rational if its ecological validity is greater than 0.5. But since this ecological validity is unknown in advance and only learnable in retrospect, this kind of selection rule is not very helpful. Moreover, the rule is incorrect inasmuch as anybody who possesses a better method than the recognition heuristic should apply this method instead of the recognition heuristic. Concerning the take-the-best heuristic TTB, Todd and Gigerenzer (2012, p. 9) assert (like Rieskamp and Dieckmann, 2012) that TTB is ecologically rational in environments with high cue redundancy and highly varied cue validities, while linear weighting-rules are said to be rational in the opposite types of environments. However, as explained in the preceding section, the connection between high cue redundancy and TTB's optimality can be violated in both directions (see Appendix); so this rule is also incorrect.

The preceding observations do not diminish the great success of the adaptive rationality program in discovering surprising "less is more" effects. They rather point toward an underdeveloped area in this program, namely the selection-of-methods problem. They also indicate a major challenge, and to a certain degree even a *dilemma*, for the program of ecological rationality. For *if there were* simple rules of the form "In environment of type E_i, method M_i is optimal" (for $i \in \{1,\ldots,n\}$), then the combined strategy "For all $i \in \{1,\ldots,n\}$: apply method M_i in environment E_i" would be a universally optimal strategy. The existence of such a strategy would, thereby, *re-install* universal rationality, and *undermine* the very program of adaptive rationality.

Can universal rationality be re-installed in this simple way? The answer is: *No*. Following from well-known results in formal learning theory (Kelly, 1996) and meta-induction (Schurz, 2008), there cannot be an inductive prediction or inference method which is optimal in *all* environments among *all* possible prediction methods. This fact has been frequently mentioned by researchers within the adaptive rationality program (cf. Todd and Gigerenzer, 2012, p. 5). A consequence of the cited result is that there cannot be exhaustive and fully general meta-rules which specify for each task and environment a locally optimal method.

[14]Moreover, Rieskamp and Dieckmann report that in environments of low redundancy, TTB performs better if the dispersion of the cues' validities is high. Logistic regression performs better than Franklin's rule for training set sizes of greater than 20%, except in environments of high redundancy and low validity dispersion, in which logistic regression beats the other methods only for training sets greater than 80%.

Schurz and Thorn (in review) call this fact the *revenge of ecological rationality*.

While there is no 'absolutely' optimal selection strategy, the ecological rationality program presupposes selection rules that are at least "very" or "sufficiently" general. Obviously, selection strategies can only have a cognitive benefit if their success is highly general, applying to a large class of environments and tasks. If such general selection strategies did not exist, one could not explain why humans are so successful in selecting the 'right' method for their given environment, in spite of the fact that their environment constantly changes.

What makes it difficult to find general rules for selecting methods is that the *success-relevant features* of the environment are frequently cognitively inaccessible. Similarly, *changes* in the environment are often unrecognizable and unforeseeable. To deal with changing environments of this sort, one needs strategies for *learning* which locally adapted methods perform best in which environment, or in which temporal phases of the environment. This brings us to the account of strategy selection by *learning* proposed by Rieskamp and Otto (2006) and the more general account of *meta-induction* developed in Schurz (2008) and Schurz and Thorn (in review). While Rieskamp and Otto suggest *reinforcement* as the learning method for strategy selection, *meta-induction* is a more general family of meta-level selection strategies which includes reinforcement as a special case.

The account of meta-induction was developed within the domain of epistemology as a means of addressing Hume's problem of induction (Schurz, 2008, 2009; Vickers, 2010), thereby utilizing certain results from the domain of machine learning (Cesa-Bianchi and Lugosi, 2006). In this account, meta-inductive selection strategies are considered as *meta-level* strategies. Such strategies attempt to select an optimal prediction method, or to construct an optimal *combination* of such methods, out of the *toolbox* of locally adapted prediction methods, which are also called the object-level methods. Meta-inductive strategies base their predictions on the *so-far observed success rates* of the available object-level methods. The simplest meta-inductive strategy is again TTB, which imitates the predictions of the so-far best available prediction method. The difference between the model envisioned here and the typical experimental paradigm used within adaptive rationality research is that within the present model TTB is applied at the meta-level, as a means to selecting the right (combination of) locally adapted prediction methods, rather than to the selection of "cues."

Recall the negative result that there is no method which is optimal among *all* possible prediction methods in all environments. In other words, no method is *absolutely* optimal. Of course, only a fraction of all possible prediction methods is cognitively accessible to any human-like agent. So at the meta-level, it is only possible to include *the cognitively accessible* prediction methods in the "toolbox" of candidate methods. This raises the following question: is there a meta-inductive strategy which predicts optimally in comparison to all candidate prediction methods that are *accessible* to it, no matter what these methods are and in which environment one happens to be? Schurz and Thorn (in review) call this property *access-optimality* (i.e., optimality among all accessible methods), in distinction to absolute optimality, which is not restricted to the accessible methods.

The philosophical importance of this notion is this: if one could prove that a universally access-optimal selection strategy exists, its application would always be reasonable, independent from one's environment and one's toolbox, because by applying this meta-strategy to the methods in one's toolbox, one can only improve but never worsen one's success rate. Arguably, the existence of such a method would also give us at least a partial solution to Hume's problem of induction (Schurz, 2008). It can easily be shown that TTB is *not* universally access-optimal: it fails to be access-optimal in environments where the success rates of the available candidate methods are constantly oscillating (Schurz, 2008, Figures 1, 4). However, there is a certain linear weighting strategy, so far unrecognized within the adaptive rationality research community, which is demonstrably universally access-optimal *in the long run*. Schurz and Thorn (in review) call this strategy *attractivity-based* weighting, AW, since it bases its assignment of weights to the predictions of accessible methods on the "attractivities" of those methods, which depend on the success differences between the object-level methods and AW. In the *short* run, AW may earn a small loss (compared to the so-far best prediction method) which vanishes if the number of rounds becomes large compared to the number of competing methods.

There are meta-level methods whose performance exceeds that of AW in particular environments. For example, Schurz and Thorn (in review, sec. 7) show that Franklin's rule (if applied at the meta-level) outperforms AW in certain environments, but is worse than AW in other environments. In other words, all improvements of AW are *local* and come at the cost of losing universal access-optimality. This is again a "revenge effect" of ecological rationality, which puts us into the following dilemma: on the one hand, there is a meta-level strategy, namely AW, which is universally access-optimal in the long run. On the other hand, there are methods whose performance may exceed that of AW locally, but only on the cost of losing universal access-optimality.

Schurz and Thorn (in review) propose to solve this dilemma by the following *division of labor*: At the meta-level of selection strategies, one should use a strategy which is access-optimal, i.e., the strategy AW. If one finds a meta-method M* which is more successful than AW in some environments, then one can improve the success of AW not by replacing AW by M* at the meta-level, but by putting M* into the toolbox of locally adapted methods and applying AW to this *extended* toolbox.

Generally speaking, the account of Schurz and Thorn (in review) proposes a division of labor between *general* meta-level selection strategies and optimal (combinations of) *locally adapted* cognitive methods. This proposed division of labor is akin to the *dual process accounts of cognition* that have been developed in the recent decades by a variety of psychologists[15].

[15]Cf. Evans and Over (1996), Sloman (1996), and Stanovich (1999); for excellent overviews cf. Evans (2003, 2008).

These accounts explain human cognition by a division of labor between two reasoning systems and corresponding processes: "type 1" processes are usually characterized as unconscious or implicit, heuristic, context-specific, perception- or action-related, fast and parallel, and evolutionarily old (humans share them with animals). In contrast, "type 2" processes are characterized as conscious and explicit, analytic, context-general, symbolic, slow and sequential, being an evolutionarily recent feature of homo sapiens.

Although the fit of the dualistic account of local methods and meta-inductive strategies with dual process accounts is not perfect, the basic similarities are clear. *Firstly*, the distinction between locally adapted prediction strategies and general (meta-inductive) selection strategies is a distinction between types of cognitive *processes* (not between cognitive "rationalities"); so it rightly belongs to the cognitive process level to which the type 1/2 distinction applies (cf. Oaksford and Chater, 2012). *Secondly*, meta-inductive strategies are conscious selection processes and thus belong to the family of type 2 processes. In contrast, locally adapted prediction or decision heuristics are often (though not always) type 1 processes, whose control by type 2 processes is difficult and requires cognitive training (Houde et al., 2000) and general intelligence (Stanovich, 1999).

The preceding short remark concerning the relation between the proposed dual account and contemporary dual process theories must be sufficient. The main purpose of the dual account of local methods and meta-strategies is to highlight the evolutionary benefit of a division of labor between general cognitive selection strategies and locally adapted cognitive methods. In particular, this division of labor helps to solve the explained dilemma facing ecological rationality program, i.e., the problem of explaining how locally adapted reasoning strategies can be cognitively successful in a situation of changing environments.

CONCLUSION

I began this article by outlining the distinction between normative and instrumental conceptions of cognitive rationality (Elqayam and Evans, 2011). The latter distinction was embedded into a broader philosophical framework, classifying the former account as deontological and the latter as teleological. While I agreed with Elqayam and Evans' critique of unjustified is-ought inferences in normative accounts, I argued that in both accounts one must make at least some value assumptions, which are based on some form of intuition. However, while those accounts which Elqayam and Evans call "normativist" are based on a mixed bag of a-priori intuitions, instrumentalist accounts are based on just one value—the value of cognitive success in the given environment. I, therefore, proposed to replace the normative/instrumental distinction by the distinction between a-priori intuition-based vs. a-posteriori success-based accounts of rationality. Cognitive success should be understood as success in finding as many relevant truths as possible with as few mistakes and cognitive costs. I argued that the value of cognitive success lies in its instrumental rationality for *almost-all* practical purposes.

After distinguishing between a-priori intuition-based vs. a-posteriori success-based accounts of rationality, I pointed out that this distinction is usually conflated with a second distinction: that between logically general vs. locally adapted rationality. In opposition to this conflation, I argued that these two distinctions should be treated as *independent* dimensions of classification. The question of whether logico-general or locally adapted reasoning methods have greater cognitive success is a descriptive question which can be decided by computational and empirical means. In the case of classical logic and probability theory, I demonstrated that logico-general systems of reasoning can be instrumentally justified by their a-posteriori cognitive success. It turns out that although reasoning according to classical logic and probability theory have the advantage of preserving truth and avoiding inconsistency in all environments, they are not cognitively successful in all situations, but only in those where they are paired with capacity for successful inductive reasoning, which supplies deductive or probabilistic reasoning with general premises about empirical regularities. In the area of inductive inference, however, there is no generally reliable or optimal reasoning method. No wonder, then, that this is the domain in which the paradigm of locally adaptive or ecological rationality has emerged.

In the final part of the paper, I argued that the fact that human beings frequently encounter changing environments generates a dilemma for the program of ecological rationality. On the one hand, this program requires rules which specify for each heuristic the kind of environment in which it may be successfully applied, and in which it may not. On the other hand, some general arguments show that a complete list of rules of this sort does not exist; in fact, its existence would undermine the very program of ecological rationality. As a way out of this dilemma, I argued for a dual account of cognition, in which highly general meta-inductive selection strategies are applied to a toolbox of locally adapted cognitive methods.

ACKNOWLEDGMENT

For valuable help I am indebted to Shira Elqayam, Jonathan Evans, P. Thorn, G. Kleiter, N. Pfeifer, R. Hertwig, and P. Pedersen.

REFERENCES

Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht: Reidel.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy* 32, 1–19. doi: 10.1017/S0031819100037943

Barbey, A. K., and Sloman, S. A. (2007). Base rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297. doi: 10.1017/S0140525X07001653

Baron, J. (2008). *Thinking and deciding*. New York, NY: Cambridge Univ. Press.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond. Biol. Sci.* 53, 370–418. doi: 10.1098/rstl.1763.0053

Birnbacher, D. (2003). *Analytische Einführung in die Ethik*. Berlin: de Gruyter.

Brighton, H., and Gigerenzer, G. (2012). "How heuristics handle uncertainty," in *Ecological Rationality: Intelligence in the World*, eds P. M. Todd and G. Gigerenzer (New York, NY: Oxford University Press), 33–60.

Broad, C. D. (1930). *Five Types of Ethical Theory*. London: Routledge.

Cesa-Bianchi, N., and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge: Cambridge Univ. Press.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behav. Brain Sci.* 4, 317–370. doi: 10.1017/S0140525X00009092

David, M. (2005). "Truth as the primary epistemic goal: a working hypothesis," in *Contemporary Debates in Epistemology*, eds M. Steup and E. Sosa (Oxford: Blackwell Publishing), 296–312.

De Finetti, B. (1937). "Foresight: its logical laws, its subjective sources," in *Studies in Subjective Probability*, eds H. E. Kyburg and H. E. Smokler (New York, NY: Wiley), 93–158.

Elqayam, S., and Evans, J. S. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–290. doi: 10.1017/S0140525X1100001X

Evans, J. S. (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychol. Bull.* 128, 978–996. doi: 10.1037/0033-2909.128.6.978

Evans, J. S. (2003). In two minds. Dual process accounts of reasoning. *Trends Cogn. Sci.* 7, 454–459. doi: 10.1016/j.tics.2003.08.012

Evans, J. S. (2008). Dual process accounts of reasoning, judgement, and social cognition. *Annu. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Evans, J. S., Handley, S. J., and Over, D. (2003). Conditionals and conditional probability. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 321–335. doi: 10.1037/0278-7393.29.2.321

Evans, J. S., and Over, D. (1996). *Rationality and Reasoning.* New York, NY: Psychology Press.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. New York, NY: Hafner Press.

Frankena, W. (1963). *Ethics*. Englweood Cliffs, NJ: Prentice Hall.

Galvan, S. (1988). Underivability results in mixed systems of monadic deontic logic. *Logique Analyse* 121, 45–68.

Gigerenzer, G. (1991). "How to make cognitive illusions disappear: beyond "heuristics" and biases," in *European Review of Psychology*, eds W. Stroebe and M. Hewstone (Chichester: Wiley), 83–115.

Gigerenzer, G., Todd, P. M., and The ABC Research Group (eds.). (1999). *Simple Heuristics That Make Us Smart*. Oxford: Oxford Univ. Press.

Gillies, D. (2000). *Philosophical Theories of Probability*. London: Routledge.

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goodman, N. (1955). *Fact, Fiction and Forecast*. Cambridge, MA: Harvard Univ. Press.

Greco, J., and Turri, J. (2013). "Virtue epistemology," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Winter 2013 Edition). Available online at: http://plato.stanford.edu/archives/win2013/entries/epistemology-virtue/

Habermas, J. (1966). Knowledge and interest. *Inquiry* 9, 285–300. doi: 10.1080/00201746608601463

Houde, O., Zago, L., Mellet, E., Montier, S., Pineau, A., Mazoyer, B., et al. (2000). Shifting from the perceptual brain to the logical brain: the neural impact of cognitive inhibition training. *J. Cogn. Neurosci.* 12, 721–728. doi: 10.1162/089892900562525

Hudson, W. D. (ed.). (1969). *The Is-Ought-Question*. London: Macmillan Press.

Hume, D. (1739/40). *A Treatise of Human Nature. Book III: Of Morals*. Mineola, NY: Dover Pub. 2004. Available online at: www.gutenberg.org/ebooks/4705

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgement of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kelly, K. (1996). *The Logic of Reliable Inquiry*. New York, NY: Oxford Univ. Press.

Moore, G. E. (1903). *Principia Ethica*. New York, NY: Cambridge Univ. Press.

Oaksford, M., and Chater, N. (1991). Against logicist cognitive science. *Mind Lang.* 6, 1–38. doi: 10.1111/j.1468-0017.1991.tb00173.x

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 603–631. doi: 10.1037/0033-295X.101.4.608

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Cambridge: Oxford University Press.

Oaksford, M., and Chater, N. (2012). Dual processes, probabilities, and cognitive architecture. *Mind Soc.* 11, 15–26. doi: 10.1007/s11299-011-0096-3

Piatelli-Palmarini, M. (1996). *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*. New York, NY: John Wiley & Sons.

Pigden, C. R. (1989). Logic and the autonomy of ethic. *Australas. J. Philos.* 67, 127–151. doi: 10.1080/00048408912343731

Pigden, C. R. (ed.). (2010). *Hume on 'Is,' and 'Ought.'* Hampshire: Palgrave Macmillan.

Ramsey, F. P. (1926). "Truth and probability," in *Philosophical Papers*, ed H. D. Mellor (Cambridge: Cambridge Univ. Press). Reprinted in Ramsey, F. P. (1990), 52–94.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard Univ. Press.

Reichenbach, H. (1949). *The Theory of Probability*. Berkeley, CA: University of California Press.

Rieskamp, J., and Otto, P. (2006). SSL: a theory of how people learn to select strategies. *J. Exp. Psychol. Gen.* 135, 207–236. doi: 10.1037/0096-3445.135.2.207

Rieskamp, J., and Dieckmann, A. (2012). "Redundancy: environment structure that simple heuristics can exploit," in *Ecological Rationality: Intelligence in the World*, eds P. M. Todd and G. Gigerenzer (New York, NY: Oxford University Press), 187–215.

Prior, A. N. (1960). The autonomy of ethics. *Australas. J. Philos.* 38, 199–206. doi: 10.1080/00048406085200221

Schurz, G. (1997). *The Is-Ought Problem: an Investigation in Philosophical Logic*. Dordrecht: Kluwer.

Schurz, G. (2001a). "Kinds of rationality and their role in evolution," in *Rationality and Irrationality*, eds B. B. Brogaard and B. Smith (Vienna: öbv & hpt), 301–310.

Schurz, G. (2001b). What is 'normal'? An evolution-theoretic foundation of normic laws and their relation to statistical normality. *Philos. Sci.* 28, 476–497. doi: 10.1086/392938

Schurz, G. (2002). Ceteris paribus laws: classification and deconstruction. *Erkenntnis* 57, 351–372. doi: 10.1023/A:1021582327947

Schurz, G. (2005). Non-monotonic reasoning from an evolutionary viewpoint: ontic, logical and cognitive foundations. *Synthese* 146, 37–51. doi: 10.1007/s11229-005-9067-8

Schurz, G. (2007). "Human conditional reasoning explained by non-monotonicity and probability," in *Proceedings of EuroCogSci07. The European Cognitive Science Conference 2007*, eds S. Vosniadou, D. Kayser, and A. Protopapas (New York, NY: Erlbaum), 628–633.

Schurz, G. (2008). The meta-inductivist's winning strategy in the prediction game: a new approach to Hume's problem. *Philos. Sci.* 75, 278–305. doi: 10.1086/592550

Schurz, G. (2009). Meta-induction and social epistemology. *Episteme* 6, 200–220. doi: 10.3366/E1742360009000641

Schurz, G. (2010). "Non-trivial versions of Hume's is-ought thesis and their presuppositions," in *Hume on "Is" and "Ought,"* ed C. R. Pigden (Hampshire: Palgrave), 198–216.

Schurz, G. (2011a). Truth-conduciveness as the primary epistemic justification of normative systems of reasoning. *Behav. Brain Sci.* 34, 266–267. doi: 10.1017/S0140525X11000537

Schurz, G. (2011b). Verisimilitude and belief revision. With a focus on the relevant element account. *Erkenntnis* 75, 203–221. doi: 10.1007/s10670-011-9291-1

Schurz, G. (2012). "Meta-induction and the problem of fundamental disagreement," in *Epistemology: contexts, values, disagreement*, eds C. Jäger and W. Löffler (Frankfurt/M.: Ontos), 343–354.

Schurz, G. (2013). *Philosophy of Science: a Unified Approach*. New York, NY: Routledge.

Schurz, G., and Thorn, P. (2012). Reward versus risk in uncertain inference: theorems and simulations. *Rev. Symb. Logic* 5, 574–612. doi: 10.1017/S1755020312000184

Schurz, G., and Werning, M. (eds.). (2009). *Reliable Knowledge and Social Epistemology*. Amsterdam: Rodopi.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 119, 3–22. doi: 10.1037/0033-2909.119.1.3

Stanovich, K. E. (1999). *Who is Rational? Studies of Individual Differences in Reasoning.* Mahwah, NJ: Erlbaum.

Stich, S. (1990). *The Fragmentation of Reason*. Cambridge, MA: MIT Press.

Stuhlmann-Laeisz, R. (1983). *Das Sein-Sollen-Problem. Eine modallogische Studie*. Stuttgart: Frommann-Holzboog.

Suppes, P. (1966). "Probabilistic inference and the concept of total evidence," in *Aspects of Inductive Logic*, eds J. Hintikka and P. Suppes (Amsterdam: North-Holland Publ. Comp.), 49–65.

Taylor, S. E. (1989). *Positive Illusions. Creative Self-Deception and the Healthy Mind*. New York, NY: Basic Books.

Thorn, P., and Schurz, G. (2014). *A Utility Based Evaluation of Logico-Probabilistic Systems*. Studia Logica. Available online at: http://link.springer.com/article/10.1007/s11225-013-9526-z

Todd, P. M., and Gigerenzer, G. (2012). "What is ecological rationality," in *Ecological Rationality: Intelligence in the World*, eds P. M. Todd and G. Gigerenzer (New York, NY: Oxford University Press), 3–30.

Unterhuber, M., and Schurz, G. (2013). The new Tweety puzzle: arguments against monistic Bayesian approaches in epistemology and cognitive science. *Synthese* 190, 1407–1435. doi: 10.1007/s11229-012-0159-y

Vickers, J. (2010). "The problem of induction," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Spring 2010 Edition). Available online at: http://plato.stanford.edu/archives/spr2010/entries/induction-problem/

Von Mises, R. (1964). *Mathematical Theory of Probability and Statistics*. New York, NY: Academic Press.

Wason, P. C. (1966). "Reasoning," in *New Horizons in Psychology*, Vol. 1, ed B.M. Foss (London: Penguin), 106–136.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 February 2014; accepted: 02 June 2014; published online: 01 July 2014.
*Citation: Schurz G (2014) Cognitive success: instrumental justifications of normative systems of reasoning. Front. Psychol. **5**:625. doi: 10.3389/fpsyg.2014.00625*
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

APPENDIX: TTB vs. FRANKLIN'S RULE IN ENVIRONMENTS OF DIFFERENT REDUNDANCY (WITH P. THORN)

Figures A1 + A2 illustrate a simulation of a prediction tournament in which Franklin's rule predicts better than TTB, independently of the cue redundancy of the environment. The value of a binary target variable (with values 1, 0) had to be predicted based on three binary cues (with values 1, 0) whose conditional success probabilities were as follows:

$$p(\text{event} = 1|\text{3-of-3 cues predict 1}) = 0.9$$
$$p(\text{event} = 1|\text{2-of-3 cues predict 1}) = 0.7$$
$$p(\text{event} = 1|\text{1-of-3 cues predict 1}) = 0.3$$
$$p(\text{event} = 1|\text{0-of-3 cues predict 1}) = 0.1.$$

The validities were assumed to be known so that learning errors play no role. By assuming different prior probabilities over the cues' combined predictions, one can make them uncorrelated (non-redundant) or highly correlated (redundant), without changing the result that in this environment Franklin's rule performs better than TTB, as shown in **Figures A1 + A2**.

Figures A3 + A4 exhibit an environment in which TTB predicts better than Franklin's rule, independently of the degree of cue redundancy. Here the success probabilities of the three cues were as follows (with "$C_{2/3}$" for "cue 2" or "cue 3"):

$$p(\text{event} = 1|C_1 \text{ predicts } 1, C_{2/3} \text{ predicts } x_{2/3} \in \{0, 1\})$$
$$= 0.9 \text{ for all choices] of } x_{2/3}.$$
$$p(\text{event} = 0|C_1 \text{ predicts } 0, C_{2/3} \text{ predicts } x_{2/3} \in \{0, 1\})$$
$$= 0.8 \text{ for all choices of } x_{2/3}.$$

By assuming either uniform prior distributions over the combined predictions or positive correlations between the cues' predictions one now obtains the result that TTB predicts better than Franklin's rule, both in low and high redundancy environments, as shown in **Figures A3 + A4**.

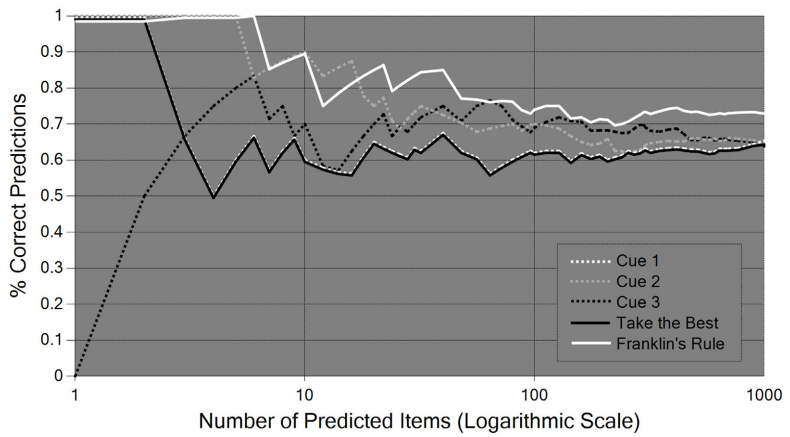


FIGURE A1

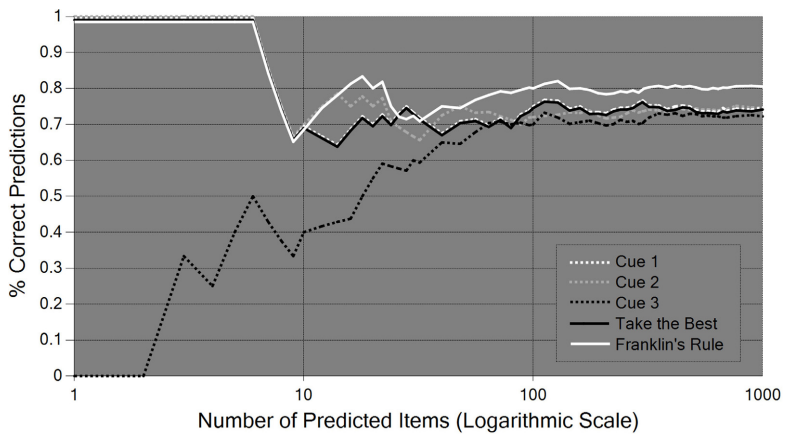


FIGURE A2

FIGURE A1 + A2 | TTB against Franklin's rule in a binary prediction tournament with known validities which are conditionally dependent. Figure A1 with low redundancy and **Figure A2** with high redundancy of cues. In both cases, Franklin's rule predicts better than TTB

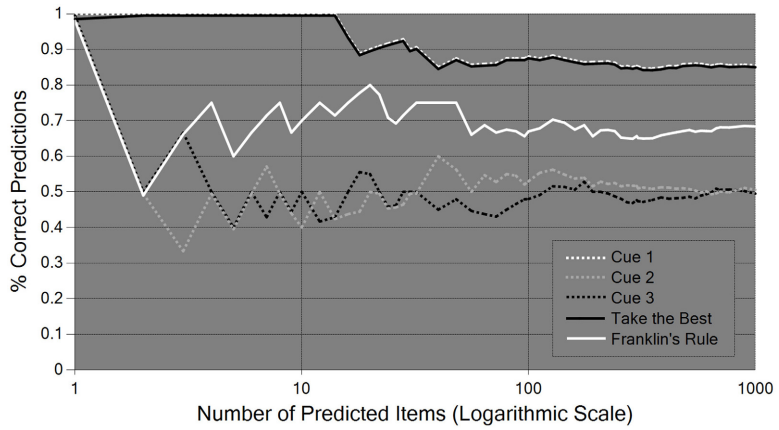


FIGURE A3

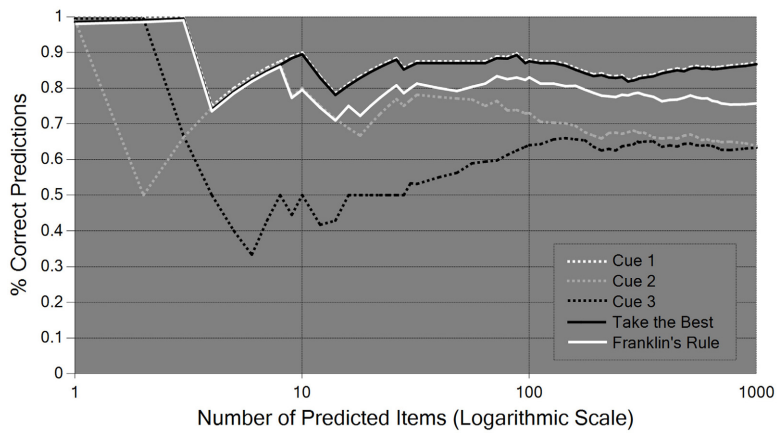


FIGURE A4

FIGURE A3 + A4 | TTB against Franklin's rule in a binary prediction tournament with known validities which are conditionally independent. Figure A3 with low redundancy and **Figure A4** with high redundancy of cues. In both cases TTB predicts better than Franklin's rule.

Empirical reports

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 17 April 2014
doi: 10.3389/fpsyg.2014.00316

New normative standards of conditional reasoning and the dual-source model

***Henrik Singmann*[1]*, *Karl Christoph Klauer*[1] *and David Over*[2]**

[1] *Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany*
[2] *Department of Psychology, Durham University, Durham, UK*

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Gordon Pennycook, University of Waterloo, Canada
Niki Pfeifer, Ludwig-Maximilians-Universität München, Germany

***Correspondence:**
Henrik Singmann, Abteilung für Sozialpsychologie und Methodenlehre, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Engelbergerstr. 41, D-79085 Freiburg, Germany
e-mail: henrik.singmann@psychologie.uni-freiburg.de

There has been a major shift in research on human reasoning toward Bayesian and probabilistic approaches, which has been called a new paradigm. The new paradigm sees most everyday and scientific reasoning as taking place in a context of uncertainty, and inference is from uncertain beliefs and not from arbitrary assumptions. In this manuscript we present an empirical test of normative standards in the new paradigm using a novel probabilized conditional reasoning task. Our results indicated that for everyday conditional with at least a weak causal connection between antecedent and consequent only the conditional probability of the consequent given antecedent contributes unique variance to predicting the probability of conditional, but not the probability of the conjunction, nor the probability of the material conditional. Regarding normative accounts of reasoning, we found significant evidence that participants' responses were confidence preserving (i.e., p-valid in the sense of Adams 1998) for MP inferences, but not for MT inferences. Additionally, only for MP inferences and to a lesser degree for DA inferences did the rate of responses inside the coherence intervals defined by mental probability logic (Pfeifer and Kleiter, 2005, 2010) exceed chance levels. In contrast to the normative accounts, the dual-source model (Klauer et al., 2010) is a descriptive model. It posits that participants integrate their background knowledge (i.e., the type of information primary to the normative approaches) and their subjective probability that a conclusion is seen as warranted based on its logical form. Model fits showed that the dual-source model, which employed participants' responses to a deductive task with abstract contents to estimate the form-based component, provided as good an account of the data as a model that solely used data from the probabilized conditional reasoning task.

Keywords: conditional reasoning, probabilistic reasoning, new paradigm psychology of reasoning, dual-source model, coherence, p-validity, rationality, mixed models

INTRODUCTION

The most influential work in the psychology of conditional reasoning long presupposed as its normative standard the binary and extensional logic of the propositional calculus (Johnson-Laird and Byrne, 1991 see especially pp. 7 and 74). In this logical system, a conditional "if p then q" is the material, truth functional conditional, which is logically equivalent to "not-p or q." There are, however, many problems with holding that the natural language conditionals that people reason with are equivalent to material conditionals (Evans and Over, 2004). Prominent among these problems are the "paradoxes" of the material conditional. For example, it is logically valid to infer a material conditional, equivalent to "not-p or q," from "not-p," and so the probability of such a conditional will increase as the probability of "not-p" increases. But consider a conditional about a coin we know to be fair, "If we spin the coin 100 times then we will get 100 heads." It would be absurd if our subjective probability for this conditional increased to ever higher levels as it became more and more likely that we would not go to the trouble of spinning the coin that many times.

Another limitation of this binary and extensional paradigm was that participants were asked in experiments on reasoning to assume that the premises were true and to give binary responses about what did, or did not, necessarily follow. In contrast, most human reasoning, in everyday affairs and science, is from uncertain premises, from more or less confidently held beliefs and statements or claims made by other people. The conclusions drawn are also more or less subjectively probable. Dissatisfaction with the traditional experiments has been a factor in the proposal of a new paradigm in the psychology of reasoning (Over, 2009; Evans, 2012; Elqayam and Over, 2013). The aim of the new paradigm is to move beyond experiments on abstract materials and premises given as assumptions. Participants are asked to reason in an everyday setting from content rich materials, and to provide their responses on graded scale reflecting various degrees of belief (see Rips, 2001; Singmann and Klauer, 2011, for an empirical dissociation of both methods).

The proposed normative system for the new studies of conditional reasoning is no longer the binary and extensional propositional calculus, but rather subjective probability theory that goes back to de Finetti (1936, 1937) and Ramsey (1931). The relevant normative standard is what de Finetti termed the logic of probability and Ramsey the logic of partial belief, as developed

by Adams (1998), Gilio (2002), Gilio and Over (2012), and others. The new paradigm in the psychology of reasoning can also be seen as part of the great impact Bayesian approaches have had generally in cognitive science (Oaksford and Chater, 1994, 2001, 2007; Oaksford et al., 2000). Evans and Over (2004), Pfeifer and Kleiter (2005, 2010), and Oaksford and Chater (2007) have proposed accounts of human conditional reasoning that are central examples of the new paradigm.

Much research in the old paradigm dealt with so-called *basic* conditionals, which are defined to be indicative conditionals with an abstract content (Johnson-Laird and Byrne, 2002). People cannot use background knowledge and context to help them evaluate basic conditionals. Such conditionals are not very similar to the knowledge and context laden conditionals of ordinary and scientific reasoning. Realistic indicative conditionals, of the latter type, can be classified in a number of ways (Douven and Verbrugge, 2010), but here we are mainly concerned with conditionals that are justified by some sort of (at least weak) causal connection between the antecedent and consequent (Over et al., 2007). For lack of a better term, we call these conditionals *everyday conditionals*. Our interest in these conditionals stems from the fact that people use subjective probability judgments based on knowledge of content and context to evaluate them.

The new paradigm gives a new interpretation to such everyday conditionals. It does not see them as material conditionals, the probability of which is the same as that the probability of "$\neg p$ or q," $P(\neg p \vee q)$, (where $\neg p$ is "not-p"). In the new paradigm, the probability of one of these conditionals, $P(\text{if } p \text{ then } q)$, is the conditional probability of its consequent given its antecedent, $P(q|p)$. The relation, $P(\text{if } p \text{ then } q) = P(q|p)$, is so important that it is simply called the Equation (Edgington, 1995) in analytical philosophy (or *conditional probability hypothesis* in psychology)[1]. Based on the Equation probabilistic accounts of human conditional reasoning were developed (Oaksford and Chater, 2007; Pfeifer and Kleiter, 2010). Moreover, if the Equation holds, the "paradoxes" of the material conditional we referred to above cannot be derived (Pfeifer, 2014; see Pfeifer, 2013, for an empirical study of the "paradoxes"). For example, it will no longer hold that the probability of the above example conditional, about spinning the coin 100 times, increases as we become more and more determined not to spin it that many times. The conditional probability that we will get 100 heads given that we spin the coin 100 times is extremely low, and will stay low as it gets more and more likely we will not spin the coin.

Another hypothesis for the probability of everyday conditionals concerns those justified by reference to causal relations. If such conditionals state the existence of a causal relation, then the presence of the antecedent should raise the probability of the consequent compared to when the antecedent is absent. In other words, whereas the conditional probability $P(q|p)$ should be positively related with the probability of a conditional, the conditional probability of alternatives to the conditional (i.e., not p cases leading to q), $P(q|\neg p)$ should be negatively related with the probability of a conditional. This inequality ($P(q|p) - P(q|\neg p) > 0$) is also known as the *delta-p* rule (Allan, 1980; Sloman, 2005).

A multitude of studies has shown that the conditional probability $P(q|p)$ and to a lesser extent the conjunction $P(p \wedge q)$ are predictors for the probability of the conditional (Evans et al., 2003; Oberauer and Wilhelm, 2003; Oberauer et al., 2007; Over et al., 2007; Douven and Verbrugge, 2010, 2013; Politzer et al., 2010; Fugard et al., 2011). A first goal of the current manuscript is to further advance these previous studies by adopting some procedural variations which more strongly capture the central notions of the new paradigm, everyday reasoning and subjective probabilities. Specifically, some of the studies (Evans et al., 2003; Oberauer et al., 2007; Fugard et al., 2011) have, using the probabilistic truth table task, provided participants with the frequencies constituting the joint probability distribution over antecedent and consequent for a given basic conditional. From this probability distribution the probabilities corresponding to the different hypotheses for the probability of the given conditional could be construed and compared with individuals' estimates of the probability of said conditional. The probabilities used were particularly easy to grasp (see especially Politzer et al., 2010), but perceptions of probabilities can be biased (e.g., Tversky and Kahneman, 1992). In some studies the conditional probability $P(q|p)$ is not reported directly by participants but calculated from their estimates of the unconditional probabilities constituting the joint probability distribution over antecedent and consequent (Over et al., 2007). As conditional probability is seen as primitive by some proponents of the new paradigm (and not defined over unconditional probabilities but given by people's use of the Ramsey test - see Evans and Over, 2004; Pfeifer and Kleiter, 2005) it may seem preferable to work with it as a primitive probability. Finally, we think it is important to show the relationship on an individual level (cf. Douven and Verbrugge, 2010, 2013). Hence, we assess the conditional probability hypothesis with everyday conditionals and assess the probabilities corresponding to the different competing hypotheses directly and independently.

In addition to the question on how individuals understand the conditional, the new paradigm also offers new ideas on how individuals reason from conditional inferences. The conditional inferences usually studied consist of the conditional as the major premise, a categorical minor premise, and a putative conclusion:

- *Modus Ponens* (MP): If p then q. p. Therefore q.
- *Modus Tollens* (MT): If p then q. Not q. Therefore not p.
- *Affirmation of the Consequent* (AC): If p then q. q. Therefore p.
- *Denial of the Antecedent* (DA): If p then q. Not p. Therefore not q.

In the next paragraphs we present major accounts for explaining reasoning from those conditional inferences within the new paradigm.

NORMATIVE ACCOUNTS

According to classical logic MP and MT are valid (i.e., truth preserving) inferences: the truth of the two premises necessarily

[1]Note that Lewis' famous *triviality arguments* (Lewis, 1976) which apparently show that the Equation is untenable on theoretical grounds, actually depend on an interpretation of conditionals that is empirically not supported. See Douven and Verbrugge (2013) for an extensive discussion and experimental results conclusively showing that Lewis' arguments do not apply.

entails the truth of the consequence. Likewise AC and DA are not valid, so the truth of the conclusion does not necessarily follow from true premises (i.e., drawing an AC or DA inference is considered a reasoning fallacy). But the new paradigm focuses generally on degrees of belief in premises. A normative view that builds upon degrees of belief is given by Adams' (1998) probability logic with the notion of probabilistic validity or p-validity, according to which inferences should be *confidence preserving*: a p-valid conclusion cannot be more *uncertain* than the premises on which it is based. Formally, uncertainty of an event p is defined as the complement of the probability of p, $U(p) = 1 - P(p)$, and for p-valid inferences the uncertainty of the conclusion cannot exceed the sum of the uncertainties of the premises whatever the probabilities of the premises and conclusion. Parallel to classical logic, MP and MT are p-valid and AC and DA are not p-valid. Hence another goal of the current manuscript is to provide a test of p-validity as a computational level account (in Marr's, 1982, sense) of human reasoning.

A stronger normative framework is proposed by Pfeifer and Kleiter's (2005; 2010) *mental probability logic*, as it derives probabilistically informative restrictions for all four inferences, MP, MT, AC, and DA. In contrast to Adams' (1998) notion that valid inferences are confidence preserving, they propose that reasoners' inferences should be probabilistically *coherent* (see de de Finetti, 1936; Coletti and Scozzafava, 2002; Gilio, 2002). Coherence here means that reasoners, when asked to estimate the probability of an event that stands in a relationship with other events for which the probabilities are known or estimated (e.g., the conclusion derived from a set of premises), should make an estimate that does not expose them to a Dutch book (i.e., that is coherent with the other probabilities according to coherence-based probability logic; see Pfeifer, 2014, for the relation to standard probability theory). Furthermore, in case not all probabilities necessary to calculate a point estimate for the desired event are available, the estimated probability should fall in the interval that is derived when the missing probabilities are allowed to range between 0 and 1. For example, in the case of MP, the two premises are (a) the conditional statement *if p then q* with probability $P(q|p)$ and (b) the minor premise p with probability $P(p)$ and the to be estimated probability is $P(q)$, for the conclusion q. According to the law of total probability the desired probability is given by:

$$\text{MP: } P(q) = P(q|p)P(p) + P(q|\neg p)\left(1 - P(p)\right) \qquad (1)$$

Note that we have exchanged the probability $P(\neg p)$ with its complement $1 - P(p)$ with the consequence that of the four terms on the right side, three are already present in the premises. The product of the premises is the first summand, $P(q|p)P(p)$, and the complement of the minor premise, $1 - P(p)$ is present in the second summand. Only the probability of alternatives to the conditional, $P(q|\neg p)$ (i.e., non-p cases in which the consequent holds), is less salient given that none of the premises concerns this probability. Assuming that $P(q|\neg p)$ can range from 0 to 1, we can substitute it with either 0 or 1 which gives us the coherence interval for MP:

$$\text{MP: } P(q) = \left[P(q|p)P(p),\, P(q|p)P(p) + \left(1 - P(p)\right)\right].$$

Pfeifer and Kleiter (2005; see also Wagner, 2004; Pfeifer and Kleiter, 2006) provide analogous intervals for the other inferences:

$$\begin{aligned}
\text{MT: } P(\neg p) &= \left[\max\left(\frac{1 - P(q|p) - P(\neg q)}{1 - P(q|p)},\right.\right. \\
&\qquad \left.\left.\frac{P(q|p) + P(\neg q) - 1}{P(q|p)}\right), 1\right] \\
\text{AC: } P(p) &= \left[0, \min\left(\frac{P(q)}{P(q|p)}, \frac{1 - P(q)}{1 - P(q|p)}\right)\right] \\
\text{DA: } P(\neg q) &= [1 - P(\neg p) - P(q|p)(1 - P(\neg p)), \\
&\qquad 1 - P(q|p)(1 - P(\neg p))]
\end{aligned}$$

One goal of the current manuscript is to test mental probability logic as a computational levels account of reasoning, by assessing whether or not participants responses fall in the intervals predicted by mental probability logic.

Another normative account of conditional reasoning stems from the proponents of Bayesian rationality, Oaksford and Chater (2007, chapter 5; Oaksford et al., 2000). Their probabilistic approach is couched within the same philosophical tradition as the aforementioned ones and also uses elementary probability theory to derive predictions but differs in one important aspect. It assumes that the presence of the minor premise sets the corresponding probability to one [e.g., $P(p) = 1$ for MP]. The assumed inferential step to derive an estimate of the conclusion is to conditionalize on the minor premise, the probability of the conclusion should equal the conditional probability of the conclusion given minor premise. For example for MP, the probability of the conclusion should equal the probability of the conditional, $P(q|p)$. Oaksford and Chater provided formulas to obtain point estimates for all four inferences. However, in the study reported in this manuscript we employed an experimental method in which the subjective probability of the minor premise need not equal 1. Therefore, we do not test the empirical adequacy of Oaksford and Chater's account as a computational level theory of reasoning. We follow Pfeifer and Kleiter (2006; 2007; 2009; 2010) and Evans et al. (2014) in studying whether people conform to p-validity and coherence in their conditional inferences when both premises are uncertain.

THE DUAL-SOURCE MODEL

A formal model for a descriptive account for probabilistic reasoning was proposed by Klauer et al. (2010), the *dual-source model*. It assumes that individuals integrate two different types of information (i.e., sources) when making an inference, background knowledge regarding the subject matter and information regarding the logical form of the inference. The background knowledge reflects individuals' subjective probability with which the conclusion follows from the premises given the individuals' knowledge about them. This part of the model is tied to the normative approaches presented so far in that the model assumes that for conditional inferences this probability is derived from a coherent probability distribution over p, q, and their complements. In fact, the published studies employing the dual-source model used

the formulas by Oaksford et al. (2000) to estimate the knowledge based component.

The theoretical expansion to the probabilistic approaches presented so far is the form-based component. It reflects individuals subjective probability with which an inference is warranted by the logical form (e.g., "How likely is the conclusion given that the inference is MP?"). The introduction of this part of the model was in part motivated by empirical findings that participants give higher estimate to a conclusion q when in addition to the minor premise p the conditional "if p then q" is also present (Liu, 2003). In other words, only conditionalizing on the minor premise, as proposed by Oaksford et al. (2000), does not seem to capture the complete data pattern. It should be noted that the form-based component is a subjective probability reflecting participants' belief in the logicality of logical forms and thereby not directly related to the actual logical status. To come to a blended reasoning conclusion the knowledge-based information, represented by parameter ξ, and the form-based information, represented by parameter τ, are integrated by the weighting parameter λ using Bayesian model averaging. The prediction of the dual-source model for a conditional C and inference x is given by

$$\lambda\{\tau(x) + (1 - \tau(x)) \times \xi(C, x)\} + (1 - \lambda)\xi(C, x) \qquad (2)$$

Note that in this formula, the knowledge parameters $\xi(C, x)$ enters the model in two places: in the knowledge-based component [the second summand which is weigthed with $(1 - \lambda)$], but also in the form-based part (weighted with λ). The rationale for the latter is that it is assumed that individuals, in cases when they are unsure of whether or not a conclusion is warranted by the logical form of the inference (i.e., in $(1 - \tau(x))$ cases), resort to their background knowledge, $\xi(C, x)$, as a fall-back position. One goal of the present manuscript is to apply the dual-source model in an experimental setup that strongly diverges from the experiments reported by Klauer et al. (2010), thereby providing convergent evidence for its usefulness. Furthermore, we use (a) a different formula to estimate the knowledge-based component which is based on the ideas of mental probability theory and uses no free parameters and (b) use a novel way to estimate the form-based component of the model which also does not rely on free parameters.

THE PRESENT EXPERIMENT

For an empirical test of the empirical adequacy of the approaches presented above it is necessary to obtain not only participants' responses to the conditional inferences, but also estimates of the probability of the premises and estimates of the hypothesized predictors for the probability of the conditional. Therefore, participants provided the probabilities necessary to test the aforementioned approaches in addition to estimating the probability of the conditional inferences. In this novel *probabilized conditional inference task*, participants were first asked for the probability of the conditional and then for the probability of the minor premise. Next we presented the conditional inference: the conditional and minor premise were presented together with participants' probability estimates for the premises. Participants were asked to estimate how likely the conclusion is given the information presented. After this, we asked for the remaining probabilities we were interested in for this specific content, such as the conditional probability $P(q|p)$ or the probability of alternatives to the conditional, $P(q|\neg p)$. To use this order invariantly, participants only worked on one inference for each conditional. In line with our goal to assess everyday reasoning we only used highly believable conditionals, as reasoning from unbelievable conditionals seems somewhat unnatural. To obtain estimates of participants' form-based components of the dual-source model participants performed a second task afterwards. They worked on a *deductive conditional inference task* with abstract materials and strong deductive instructions (see Singmann and Klauer, 2011).

METHODS

PARTICIPANTS

Thirty participants (mean age = 22.4 years, $SD = 2.9$, range from 18 to 30 years) participated in this experiment which was the second session in a larger study on reasoning addressing other hypotheses with other materials. In the previous session participants had worked on a conditional inference task with probabilistic instructions. More specifically, in the previous session participants were asked to provide estimates for the probability of the conclusions of all four conditional inferences (plus for the four so-called *converse inferences*; Oaksford et al., 2000) for six different conditionals three of which were uttered by an expert and three by a non–expert (i.e., analogous to Stevenson and Over, 2001). Sessions were separated by at least 1 week. Most participants were students of the University of Freiburg (28) with differing majors, excluding majors with an education in logic such as math, physics or psychology. Participants received 14€ compensation after the third session.

MATERIALS

All materials were presented in German, participants' mother tongue. For the probabilized conditional reasoning task we adapted 13 believable conditionals from Evans et al. (2010) and added three similar conditionals (which were not pretested) such as "If Greece leaves the Euro then Italy will too." The full list of conditionals can be found in the Supplemental material. Each participant worked on four randomly selected conditionals of the total of 16 conditionals and performed only one inference (i.e., MP, MT, AC, or DA) per conditional. More details are given below. In the instructions it was clarified that the conditionals were related to events that might occur within the next ten years in Germany or the rest of the world.

For the deductive conditional inference task we used two conditionals about a hypothetical letter number pair: "If the letter is a B then the number is a 7." and "If the number is a 4 then the letter is an E". Participants performed all four inferences for both conditionals.

PROCEDURE

Probabilized conditional inference task

In the first part of the experiment, participants were instructed to estimate probabilities of events or statements or to estimate the probability of a conclusion following an argument, "as if

they were in a discussion regarding these issues." Four conditionals were randomly selected for each participant and randomly assigned to the four inferences. For each conditional/inference participants responded to eight items which were presented in one block (i.e., participant first responded to all eight items for the conditional that was randomly selected for e.g., MP, before working on the eight items for the conditional that was randomly selected for e.g., MT). As participants only worked on exactly one conditional for each inference, participants worked on four blocks of eight items in total (i.e., 32 items overall in the probabilized conditional inference task) and the order of blocks was also randomized anew for each participant. For each item, the response was given on a scale from 0 to 100%. In contrast to the work of Pfeifer and Kleiter (e.g., 2007, 2010), who asked participants to provide either point estimates or interval estimates, participants in our experiments always had to provide point estimates, even for the conditional inference [type (b) below]. The responses were transformed to a probability scale (i.e., divided by 100) prior to the analysis. Each item appeared on its own screen.

Within each block, participants responded to three different types of items: (a) first participants gave estimates for the *probability of the premises*, (b) then participants had to estimate the *conclusion of the conditional inference*, and (c) finally participants had to estimate the *other probabilities* we were interested in. The three different types of items were always presented in that order. In the following we present one example for each of the eight items using the conditional "If Greece leaves the Euro then Italy will too", assuming it was randomly selected for the MP inference. For items of type (a) (probability of the premises), participants first estimated the probability of the conditional, P(if p then q), and then the probability of the minor premise, $P(p)$.

If Greece leaves the Euro then Italy will too.
In your opinion, how probable is the above statement/assertion [Aussage]?
Greece will leave the Euro.
In your opinion, how probable is it that the above event occurs [dass die obige Aussage eintritt]?

There was only one item of type (b): Participants had to give an estimate of the probability of the conclusion following the conditional inference. They, were again presented the conditional and the minor premise along with the probability estimates participants had just given (represented by *xx*):

If Greece leaves the Euro then Italy will too.
(Probability *xx*%)
Greece will leave the Euro.
(Probability *xx*%)
Under these premises, how probable is that Italy will leave the Euro, too?

After this, the items of type (c) for the other probabilities we were interested in were presented in a new random order for each block and participant. For evaluating the conditional probability hypothesis, we asked participants to estimate the conditional probability $P(q|p)$, the probability of the conjunction $P(p \wedge q)$, and the probability of the material conditional $P(\neg p \vee q)$. Furthermore we asked for the probability of alternatives, $P(q|\neg p)$, and again for the probability of the event in the conclusion this time without the premises (i.e., $P(q)$ for MP, $P(\neg p)$ for MT, $P(p)$ for AC, and $P(\neg q)$ for DA; however, we do not report an analysis of this estimate in the following):

$P(q|p)$:
How probable is that Italy will leave the Euro should Greece leave the Euro?
$P(p \wedge q)$:
Greece will leave the Euro and simultaneously Italy will leave the Euro.
In your opinion, how probable is it that the above event occurs?
$P(\neg p \vee q)$:
Greece will NOT leave the Euro or Italy will leave the Euro.
In your opinion, how probable is it that the above event occurs?
$P(q|\neg p)$:
How probable is that Italy will leave the Euro should Greece NOT leave the Euro?
$P(q)$:
Italy will leave the Euro.
In your opinion, how probable is it that the above event occurs?

After working on all eight items for one combination of inference and conditional, participants then worked on the next block of eight items with a different combination of inference and conditional. Note that in blocks for inferences other than MP, the questions for the minor premise [type (a)] and the question for the conclusion [type (b) and type (c), last item] were adapted accordingly (but no other questions).

Deductive conditional inference task

Directly after the first task the second task started, which was modeled after Singmann and Klauer's (2011) deductive condition. Participants were instructed to judge the logical validity of arguments: "Which conclusion follows with logical necessity from a given argument?" The response had to be given on a scale from 0 to 100 (i.e., the same scale as in the probabilized task, but without the %-character). For example:

If the number is a 4 then the letter is an E.
The number is a 4.
How valid is the conclusion that the letter is an E from a logical perspective?

Participants were instructed to respond with 0 if the conclusion did not necessarily follow from the premises and with 100 if the conclusion did necessarily follow from the premises. Furthermore they read: "When you are unsure, you can indicate the degree to which you think the conclusion is valid by selecting a number between 0 and 100." Participants worked on all four inferences for each of the two conditionals. Presentations of inferences was random, blocked per conditionals, with the blocks also presented

in random order. The responses were transformed to a probability scale (i.e., divided by 100) prior to the analysis.

RESULTS AND DISCUSSION

CONDITIONAL PROBABILITY HYPOTHESIS

The conditional probability hypothesis states that the probability of the conditional $P(\text{if } p \text{ then } q)$ is predicted by the conditional probability $P(q|p)$, whereas neither the probability of the material conditional $P(\neg p \vee q)$ nor the probability of the conjunction $P(p \wedge q)$ should contribute unique variance to this prediction. According to the delta-p rule, the probability of the conditional should also be negatively related to the probability of alternatives $P(q|\neg p)$. **Table 1** displays the correlations of these variables across all responses (i.e., item by participant combinations). It can be seen that, as predicted, the conditional probability $P(q|p)$ and additionally the conjunction $P(p \wedge q)$ are correlated with $P(\text{if } p \text{ then } q)$ but not the other variables. However, these results have to be interpreted cautiously as responses were nested within participants (each participant gave four responses) and within conditionals (for each conditional we obtained between five and ten responses) which violates the assumptions for standard correlation or multiple regression (Judd et al., 2012).

To overcome these problems, we estimated a linear mixed model (LMM) for the probability of the conditional as dependent variable with crossed random effects for participants and conditional (Baayen et al., 2008) using `lme4` (Bates et al., 2013) for the statistical programming language `R` (R Core Team, 2013). We entered the four assumed predictors and inference (MP, MT, AC, and DA) simultaneously as fixed effects and estimated random intercepts for participants and items plus random inference slopes and correlations among the random inference slopes for the random item effect. This model realized the *maximal random effects structure* recommended by Barr et al. (Barr, 2013; Barr et al., 2013, the random inference slopes for participants had only one observation for every level and could therefore not be estimated reliably)[2]. A model without the fixed and random effects for inference produced the exact same pattern of significant and non–significant results. To assess the significance of fixed effects in LMMs we obtained the Kenward-Rogers approximation for degrees of freedom of the full model compared with a model in which the effect of interest was excluded throughout this manuscript with the methods implemented in `afex` (Singmann, 2013) and `pbkrtest` (Halekoh and Højsgaard, 2013). The fixed effects are displayed in **Table 2** and were fully in line with the conditional probability hypothesis: when controlling for participant and item effects and estimating all parameters simultaneously, only the conditional probability $P(q|p)$ was a significant predictor of the probability of the conditional and none of the other variables. In fact, for all other predictors the estimated parameters were virtually 0.

In an exploratory analysis we estimated a second mixed model in which we added all interactions of the predictors of interest (after centering all predictors and the dependent variable on 0). The random effects structure remained identical to the previous model. In an additional exploratory analysis in which we excluded the random and fixed effects for inference, the pattern of significant and non–significant effects was the same as reported below. The analysis revealed, in addition to the significant main effect of $P(q|p)$, a significant three-way interaction of $P(q|p)$ with $P(p \wedge q)$ and $P(\neg p \vee q)$, $F_{(1,\ 72.74)} = 4.09$, $p = 0.047$ (the full results table can be found in the Supplemental material). This interaction is displayed in **Figure 1**, with the main predictor $P(q|p)$ on the x-axis and the dependent variable $P(\text{if } p \text{ then } q)$ on the y-axis, high and low values of $P(p \wedge q)$ are displayed as separate lines and high and low values of $P(\neg p \vee q)$ are displayed as separate plots (with high and low values referring to values plus and minus one SD from the mean, Cohen et al., 2002). The mean values are displayed as black lines and the individual estimates based on the random participant intercepts are displayed as gray lines in the background. Predictions were obtained by setting $P(q|\neg p)$ to 0, aggregating across all four inferences, and then transforming the predictions back on the probability scale. This interaction indicated that for low values of $P(\neg p \vee q)$, higher values for $P(p \wedge q)$ also meant higher values for $P(\text{if } p \text{ then } q)$, whereas for high values of $P(\neg p \vee q)$, $P(p \wedge q)$ interacted with $P(q|p)$ so that for

Table 1 | Correlations with the Probability of the Conditional *P*(if *p* then *q*).

	***P*(*q*|*p*)**	***P*(*p* ∧ *q*)**	***P*(¬*p* ∨ *q*)**	***P*(*q*|¬*p*)**	**Mean**	***SD***
P(if *p* then *q*)	**0.84***	**0.61***	0.09	0.04	0.61	0.26
P(*q*|*p*)		**0.72***	0.11	0.08	0.60	0.27
P(*p* ∧ *q*)			0.15	**0.21**	0.54	0.30
P(¬*p* ∨ *q*)				**0.43***	0.42	0.25
P(*q*|¬*p*)					0.27	0.24

Significant correlations ($p < 0.05$) are printed in bold. Correlations that are also significant after controlling for multiple testing using the Bonferroni-Holm correction are additionally marked with an asterisk. The two rightmost columns show mean and SD of the variables.

[2]Throughout this manuscript whenever estimating random slopes we also estimated the correlation between the slopes.

Table 2 | Main effects linear mixed model on the probability of the conditional *P*(if *p* then *q*).

Effect	**Parameter**	***F***	***df***	**F-scaling**	***p***
(Intercept)	0.14	8.70	1, 60.35	1	0.005
Inference		0.43	3, 10.16	0.84	0.74
P(*q*|*p*)	0.78	86.81	1, 88.14	1	<0.001
P(*p* ∧ *q*)	0.00	0.00	1, 90.91	1	>0.99
P(¬*p* ∨ *q*)	−0.01	0.01	1, 88.26	1	0.91
P(*q*|¬*p*)	−0.00	0.00	1, 81.59	1	0.98

The model was fitted with restricted maximum likelihood. Model df = 20, AIC = −98.42, BIC = −42.67, deviance = −138.42, Ω_0^2=0.84 (explained variance against the intercept only model; Xu, 2003). The values in the table note are based on a model with variables centered on 0 to be comparable to the Supplemental material.

high values of $P(q|p)$ lower values of $P(p \wedge q)$ predicted higher $P(\text{if } p \text{ then } q)$.

In summary, our data corroborated the conditional probability hypothesis: in contrast to previous work (e.g., Over et al., 2007; Fugard et al., 2011) only the conditional probability $P(q|p)$ is a significant predictor of the probability of the conditional. There was no evidence in support of the other hypotheses. Although we found an unexpected three-way interaction involving $P(p \wedge q)$ and $P(\neg p \vee q)$, the interaction is not easy to interpret and without proper replication we refrain from discussing it further. Our second mixed model analysis revealed another interesting finding: there does not seem to be any influence of alternatives to the conditional $P(q|\neg p)$. If the delta-p rule influenced the subjective probability of a conditional we would expect to find either a main effect or an interaction of $P(q|\neg p)$ with $P(q|p)$. As neither of those appeared delta-p is not supported by our data.

p-VALIDITY

According to Adams (1998), the p-valid inferences MP and MT are confidence preserving: the uncertainty of the conclusion should not exceed the summed uncertainty of the premises, where uncertainty is defined as $U(p) = 1 - P(p)$. **Figure 2** displays the summed uncertainties of the premises on the x-axis against the uncertainty of the conclusion on the y-axis, for the individual responses to the four inferences (summed uncertainties larger than 1 are truncated at 1). Values in the lower triangle of each panel are consistent with p-validity and values in the upper triangle can be considered violations of p-validity (this only refers to MP and MT as there is no restriction for AC and DA). The numbers in the upper left corner of each plot are the percentage of data points in the upper triangle (i.e., violations for MP and MT). Inspection of the figure reveals that there are no violations of p-validity for the forward inference MP, but there are 20% violations for the backward inference MT. One interesting finding emerges when looking at the two inferences that are not restricted by p-validity: They mimic the pattern found for MP and MT. For the other forward inference, DA, there are also no responses in the upper triangle, whereas for the other backward inference, AC, 17% of the responses are also in the upper triangle.

The analysis so far did not take into account that the larger the summed uncertainty of the premises, the larger the probability that the response to the conditional inference is p-valid (i.e., in the lower triangle) just by chance. In the extreme case of summed uncertainties of 1 (e.g., if the probabilities of the premises are .5 each or lower) the probability of giving a p-valid response is also 1. In this case, participants cannot give a response that is not p-valid, because every possible response is. When assuming that for a chance response any value is equally likely (i.e., responses are uniformly distributed across the response scale), one can control for this chance factor in the following way, as suggested by Jonathan Evans and colleagues (Evans et al., 2014). We computed a binary variable of whether or not a given response is p-valid (coded with 1) or not (coded with 0) and compared it with the sum of the uncertainties of the premises (truncated at 1), as this gives the probability of giving a p-valid response by chance. If the difference of these two variables would be above 0, the rate of responses being p-valid would be larger than expected

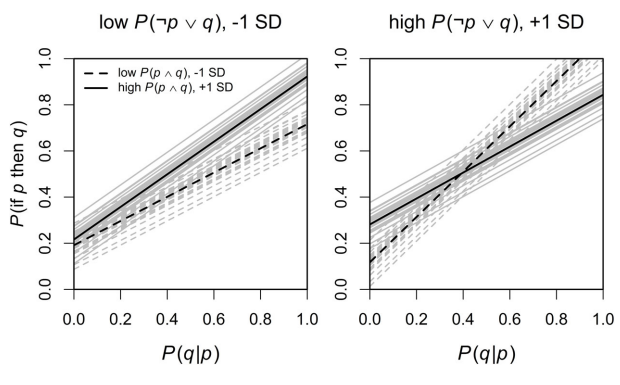


FIGURE 1 | The significant three-way interaction of $P(q|p) \times P(p \wedge q) \times P(\neg p \vee q)$ on $P(\text{if } p \text{ then } q)$. Predictions based on the random participant effects are shown in gray, and the mean effect in black. Detailed description in the main text.

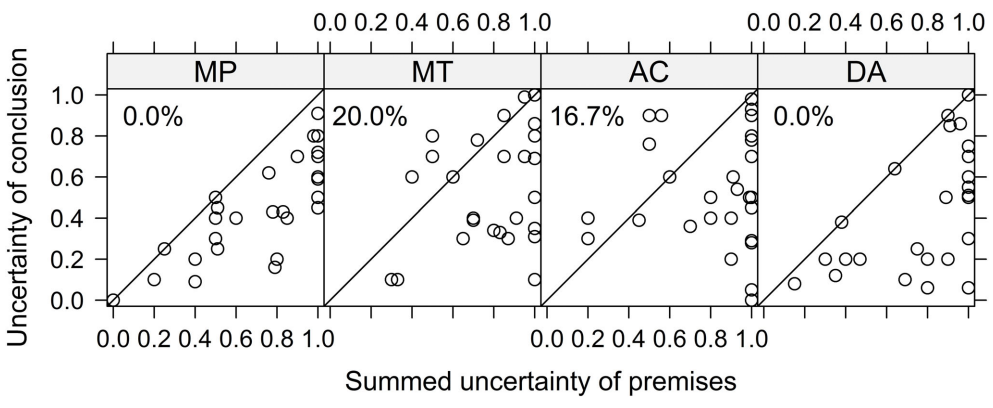


FIGURE 2 | Uncertainty of conclusion (x-axis) vs. summed uncertainties of premises (y-axis) for the four conditional inferences. Each data point shows a single response to one inference. For MP and MT, *p*-validity is violated if data points fall in the upper triangle (i.e., uncertainty of the conclusion is larger than the uncertainty of the premises). For AC and DA no such restriction exists. The numbers in the upper left corner show the percentages of data points in the upper triangle.

by uniformly distributed random responses and thus it would constitute evidence for above chance p-valid responses.

Therefore we estimated a LMM with this difference score as dependent variable with inference (MP vs. MT) as fixed effects and random intercepts for participants plus random intercepts and random inference slopes for items. The analysis showed that overall the intercept was significant, $F_{(1,\ 10.48)} = 8.39$, $p = .02$, indicating that there was evidence for above chance performance. However, the effect of inference was also significant, $F_{(1,\ 28.98)} = 8.41$, $p = 0.007$, indicating that the inferences differed in their degree of over chance performance. In fact, *post-hoc* analysis using the methods implemented in `multcomp` (Bretz et al., 2011) revealed that only for MP was the estimated effect of 0.26 reliably above zero, $z = 4.21$, $p < 0.001$. In contrast, for MT, the effect was estimated to be virtually 0 (-0.004) and consequently not significant, $z = -0.06$, $p = 0.52$. In this *post-hoc* analysis we used directional (i.e., one sided) hypotheses and the Bonferroni-Holm correction to control for alpha error cumulation.[3] As some of the violations of p-validity seemed to be rather mild violations (i.e., relatively near to the diagonal of **Figure 2**), we repeated the reported analysis after adding 0.05 and then again after adding another 0.05 (i.e., .1 in total) to the summed uncertainty of the conclusion to take minor deviations into account. These two alternative analyses yielded the exact same pattern of significant and non–significant results.

Taken together, this analysis shows that for MP, participants give p-valid inferences. In contrast, for MT individuals do not strictly draw p-valid conclusions, but sometimes are more uncertain about the conclusions than implied by the premises. Although some of those violations appear to be only mild violations (i.e., the problematic data points are near the diagonal) the analysis that takes chance into account indicates that there is overall no evidence for p-validity above chance for MT. This difference between MP and MT resembles the well-known asymmetry found in conditional reasoning with deductive instructions that individuals are more likely to endorse MP than MT inferences (e.g., Schroyens and Schaeken, 2003, Figure 4).

COHERENCE

In the next analysis we calculated coherence intervals based on mental probability logic (Pfeifer and Kleiter, 2005, 2010) for each individual response using the probability estimates of the premises. The intervals and the corresponding responses given to the conditional inferences are displayed in **Figure 3**. From this figure it is apparent that not all responses are coherent (i.e., fall within the coherence interval), however, some of those violations are very near the interval borders. Similar to p-validity, responses can fall within the intervals predicted by mental probability logic simply by chance (i.e., the larger the interval, the larger the chance to give a response within the interval). Therefore, we first looked at the correlations of the size of the interval with whether or not a response is coherent, which are given in the header of each panel in **Figure 3**. For MP there is clearly no such relationship. There is slight evidence for this correlation for MT and a clear correlation for both AC and DA.

[3] An alternative analysis comparing the observed rate of p-valid responses and the chance rate of p-valid responses using either a paired *t* test or a paired permutation test (i.e., stratified by participant) based on 100,000 Monte Carlo samples as implemented in package `coin` (Hothorn et al., 2006, 2008) yielded the same pattern of significant and non–significant results. However, as this analysis did not take potential effects of the conditionals into account (see e.g., Judd et al., 2012) we prefer to report the LMM analysis.

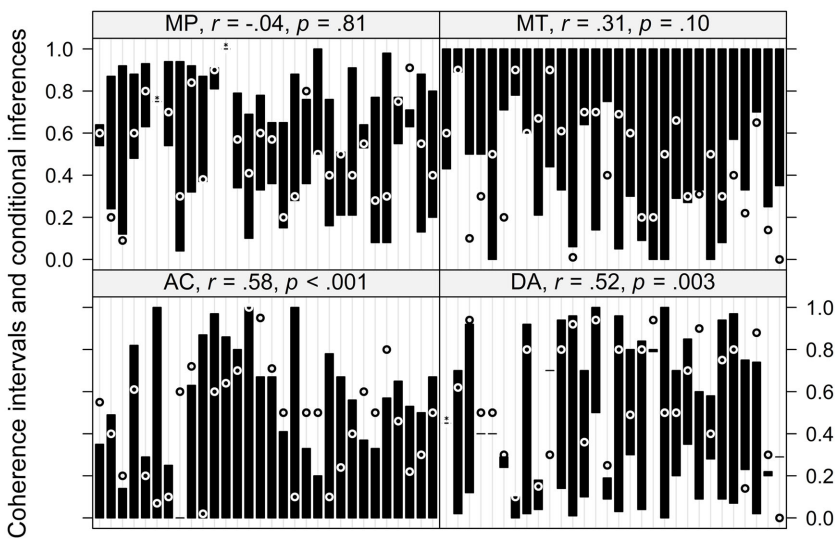


FIGURE 3 | Individual coherence intervals as predicted by mental probability logic and corresponding responses to the conditional inferences. The intervals are depicted by black bars, responses inside the interval are depicted as a white "○," responses outside the interval are depicted as a black "○." Three cases in which the interval was only of length 0.01 but the response inside the interval are marked with an asterisk. The correlation depicted in the header of each panel is the correlation of the size of each interval with whether or not a response falls within the interval. Within each panel, the x-axis is ordered by participant ID.

Next, we performed an analysis similar to the one reported for p-validity (again following Evans et al., 2014). For each participant and response we calculated whether or not a response falls within the interval or not (coded as 1 or 0, respectively) and compared it with the size of the interval as the chance level to give a response within the interval (again assuming that random responses are uniformly distributed across the response scale). These values (as percentages) are given in **Table 3**, which also contains those percentages for intervals that are extended by 0.05 or 0.1 beyond the coherence intervals. To assess if observed rates of coherent responses were larger than the chance rate of coherent responses, we estimated a LMM with the difference between both variables as dependent variable with inference (MP, MT, AC, vs. DA) as fixed effect and random intercepts for participants plus random intercepts and random inference slopes for items. The analysis revealed a significant intercept, $F_{(1,\ 16.07)} = 7.37$, $p = 0.02$, indicating above chance performance, and a marginally significant effect of inference, $F_{(3,\ 9.26)} = 2.88$, $p = 0.09$. A *post-hoc* analysis analogous to the one reported above revealed that only MP showed a significant above chance performance of 0.40, $z = 4.14$, $p < 0.001$. The only other effect that was not estimated to be virtually 0 was DA with 0.14. However, this effect did not reach significance, $z = 1.61$, $p = 0.16$ (this effect almost reached significance, $p = 0.053$, when not controlling for alpha error cumulation). The effects for MT and AC (-0.02 and 0.02, respectively), did not differ from zero, $z = -0.21$, $p = 0.82$ and $z = 0.22$, $p = 0.82$, respectively. When repeating this analysis with the extended intervals the pattern of significant and non–significant results stayed basically the same, with the only exception that for the extended intervals, the *p*-values for the effect of DA dropped below 0.05 even when controlling for alpha error cumulation.

Our analysis of the predictions of mental probability logic reveals that, similar to p-validity, participants do not strictly adhere to coherence. In fact, only for MP and to a lesser degree for DA do we find above chance performance. In addition, it should be noted that **Table 3** shows that the distance of incoherent responses from the border of the intervals is relatively large, at least for MT and AC, indicating that these outside responses are clear violations.

THE DUAL-SOURCE MODEL

Deductive conditional inference task

To fit the dual-source model to the data we combined estimates from the probabilized conditional inference task which provided the basis for the knowledge-based component of the dual-source model (more below) with the deductive conditional inference task which provided estimates for the form-based component of the dual-source model. In the latter task, we expected participants to display a pattern of results that would be consistent with what is usually found in experiments with deductive instructions and basic conditionals (e.g., Evans, 1993): Almost unanimous endorsement of MP, lower endorsement of MT, and still lower endorsement of AC and DA, with the latter two not necessarily differing. This expected pattern is essentially what we found, as evident from **Figure 4** and an LMM on the responses with inference and conditional and their interaction as fixed effects and random intercepts for participant plus random slopes for inference and conditional. We only found a significant effect of inference, $F_{(3,\ 27)} = 9.58$, $p < 0.001$, other $F < 1$. Planned comparisons using `multcomp` (Bretz et al., 2011) with directional hypotheses and no alpha-error correction revealed that indeed, endorsement for MP was higher than for MT, $z = 2.87$, $p = 0.002$, and endorsement for MT tended to be higher than endorsement for AC and DA, $z = 1.36$, $p = 0.09$, whereas there were no differences between AC and DA, $z = -0.38$, $p = 0.65$.

Specifying the model(s)

As already mentioned in the introduction, our method to estimate the dual-source model diverged from the parametrization used by Klauer et al. (2010). In particular, similar to Klauer et al. we assumed that participants' estimates of the probability of the conclusion from their background knowledge should follow from a coherent joint probability distribution over p, q and their complements. But in contrast to the original formalization which was based on Oaksford et al. (2000), we here follow the formalization of mental probability logic (Pfeifer and Kleiter, 2005, 2010) in that we assume that the law of total probability (as expressed in Equation 1 for MP) is the appropriate formula to describe this component (the corresponding formulas for the

Table 3 | Percentage of coherent responses/coherent responses predicted by chance.

interval	MP	MT	AC	DA
+/− 0	87%/45% (7%, 0.03; 7%, 0.12)	63%/65% (37%, 0.17)	60%/58% (40%, 0.18)	60%/46% (10%, 0.29; 30%, 0.10)
+/− 0.05	97%/54% (0%, 0; 3%, 0.15)	73%/69% (27%, 0.22)	63%/62% (37%, 0.15)	67%/54% (10%, 0.24; 23%, 0.05)
+/− 0.1	97%/63% (0%, 0; 3%, 0.10)	73%/73% (27%, 0.18)	73%/67% (27%, 0.12)	83%/61% (7%, 0.24; 10%, 0.04)

Percentages of responses within the coherence intervals/percentages of responses within the coherence intervals predicted from the size of the intervals. The numbers in parentheses below each row are the percentages of responses below and above the interval (only below for MT and only above for AC) and the median distance of the outside responses from the border of the interval. Rows 2, 3 present the same information but with intervals extended on both sides by 0.05 and 0.1, respectively.

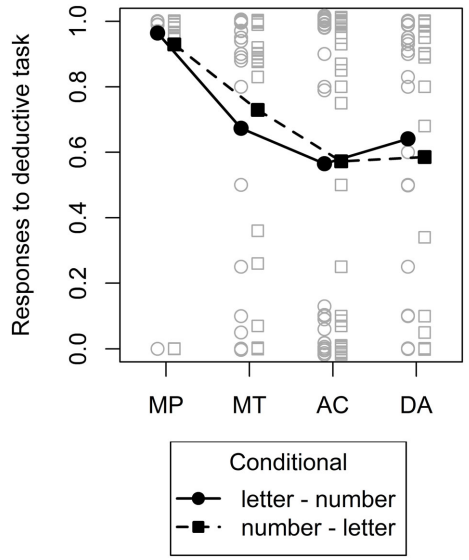


FIGURE 4 | Mean (filled symbols) and individual (non–filled symbols) responses to the deductive conditional inference task. The two different conditionals are depicted by different lines and symbols. A small amount of vertical jitter was added to individual responses in case of perfect overlap to make points distinguishable.

other inferences can be construed by elementary algebra[4]). Note that Klauer et al.'s task and the presented probabilized conditional inference task differ in that the minor premise was presented as certain in the former case and as uncertain here. The difference to the coherence intervals proposed by Pfeifer and Kleiter is that we use participants' estimate of the probability of alternatives to the conditional, $P(q|\neg p)$, which they provided after making the conditional inference, to obtain point estimates of the knowledge-based component.

The *baseline model* (BL) we compare the dual-source model against, only uses this point estimate and therefore has no free parameters. This model reflects the idea the normative accounts discussed in the introduction share that responses to conditional inferences should come from a coherent probability distribution over the elementary propositions in the inference. We use three estimates from the participants to obtain a prediction for each of the four responses to the conditional inferences: The two estimates of the premises (identical to what is used for obtaining the coherence intervals), which are obtained prior to making the conditional inference, plus the estimate of the alternatives to the conditional, $P(q|\neg p)$, which is obtained after making the conditional inference.

For estimating the *dual-source model* (DS), we combined the estimate of the baseline model as knowledge-based component of the dual source model [i.e., $\xi(C, x)$ in Equation 2] with estimates for the form based component [i.e., $\tau(x)$ in Equation 2]. As estimates of the form-based components we used participants' responses to the deductive conditional inference task (aggregating across the two different conditionals). These two types of information were integrated using the weighting parameter λ, which we treated as a free parameter (constrained to vary between 0 and 1). In sum, we used four estimates from the participants to obtain predictions for each of the four responses to the conditional inferences (i.e., the three estimates used for the baseline model plus the estimate for the corresponding inference from the deductive task) plus one free parameter per participant.

As the dual-source model now necessarily has to provide at least an as good account of the data as the baseline model (although it uses additional data, the free parameter can only increase the goodness of fit), we considered a variant of the baseline model, denoted BL*, which also included one free parameter per participant. Specifically, we wanted to acknowledge the fact that the estimate of alternatives to the conditional, $P(q|\neg p)$, was obtained after making the conditional inference. It may well be possible that participants show a bias due to memory or reevaluation effects when giving their estimates of $P(q|\neg p)$. Hence, for BL* we estimated one free parameter per participants that was multiplied with all four estimates of $P(q|\neg p)$ for that participant and could range between 0 and infinity, therefore acting as a scaling parameter for all four $P(q|\neg p)$.

We fitted all three models (BL, DS, and BL*) to the data of individual participants (i.e., to the four responses given to the four conditional inferences) using the estimates and parameters described above and using root mean squared deviation (RMSD) of predicted and observed values as criterion. For four data points from four different participants we could not obtain a prediction from the baseline model as a denominator in the formulas given in Footnote 4 was 0. We excluded these four participants from the following analysis.

Modeling results

The results from the different models as well as the original responses are displayed in **Figure 5**, the corresponding mean RMSDs are given in the lower right of the figure. To analyze the results we estimated a LMM on the individual RMSDs with model (baseline, dual-source, and BL*) as fixed effect and random intercepts for participants (random slopes for model could not be estimated as our design contained no replicates, Barr, 2013). As expected, we found a significant effect of model, $F_{(2,\ 50)} = 9.79$, $p < 0.001$. *Post-hoc* tests using Bonferroni-Holm correction for multiple comparisons revealed that, trivially, the models with a free parameter provided a better account than the baseline model, $z = -4.17$, $p = 0.003$. However, there were no differences between the dual-source model and the BL* model, $z = 1.48$, $p = 0.14$.

According to the dual-source model, individuals integrate different types of information when making a conditional inference.

[4]For MP we used Equation 1. For the other inferences we used the following formulas:

$$\text{MT: } P(\neg p) = \frac{1 - P(q|p) - P(\neg q)}{P(q|\neg p) - P(q|p)}$$

$$\text{AC: } P(p) = \frac{P(q) - P(q|\neg p)}{P(q|p) - P(q|\neg p)}$$

$$\text{DA: } P(\neg q) = 1 - P(\neg p) \times P(q|\neg p) - P(q|p)(1 - P(\neg p))$$

Values outside the probability scale (i.e., outside the interval from 0 to 1) were set to the corresponding border.

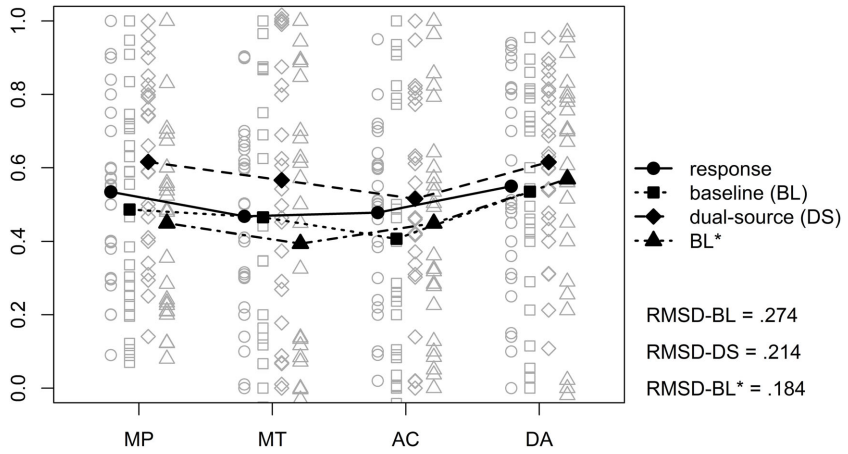


FIGURE 5 | Mean (filled symbols) and individual (non–filled symbols) model predictions and observed responses from the probabilized conditional inference task. The solid line shows the observed responses and the different dashed lines show the different models. A small amount of vertical jitter was added to individual data points in case of perfect overlap to make points distinguishable. The values in the lower right corner are the mean root mean squared deviations (RMSD) of the different models vs. the observed responses.

An analysis of the estimated λ parameters showed that 81% of the participants used the form-based information (i.e., $\lambda > 0$), which ranged for those participants from 0.02 to 0.70 with a mean of 0.31.

An analysis of the free parameter of the BL* model (i.e., the scaling parameter for all $P(q|\neg p)$ per participant) indicated that approximately half of the participants (54%) produced too large estimates of $P(q|\neg p)$, as indicated by scaling parameters below 1. The median scaling parameter was 0.95 (mean = 1.13, $sd = 1.18$). For three individuals the scaling parameter was even virtually 0, indicating that they did not consider $P(q|\neg p)$ at all in their responses to the conditional inferences. The maximal value of the scaling parameter was 6.00.

GENERAL DISCUSSION

The goals of this manuscript were to test several central assumptions of what has been introduced as the "new paradigm psychology of reasoning" (Over, 2009). The first question was how individuals understand the conditional. Specifically, we provided another test of the conditional probability hypothesis, addressing the question what predicts the probability of a conditional "If p then q," avoiding some limitations of previous assessments. Our results could not be clearer. The data supports the conditional probability hypothesis but none of the alternative explanations. Only $P(q|p)$ adds unique variance to the prediction of $P(\text{if } p \text{ then } q)$. Interestingly, another hypothesis that is associated with the new paradigm but can also be related to causal Bayes nets (e.g., Fernbach and Erb, 2013; Rottman and Hastie, 2014), the delta-p rule, receives essentially no empirical support. This is especially surprising as Douven and Verbrugge (2012) found an effect of a measure similar to delta-p when participants were asked to estimate the acceptability instead of the probability of a conditional. Furthermore, our results extend findings that there is hardly any support for the hypothesis that the conjunction $P(p \wedge q)$ predicts the probability of everyday conditionals (e.g., Over et al., 2007; Douven and Verbrugge, 2013). It seems that this latter hypothesis can only be confirmed for basic conditionals and if participants are not used to the task (Fugard et al., 2011). All in all this shows that for probabilistic tasks as employed here, the Equation offers the only supported explanation as to how participants understand a conditional. If and how causal considerations might also influence this understanding still needs to be shown.

The second main goal was to assess whether two normative accounts that have received special attention within the new paradigm, Adams' (1998) notion of p-validity and Pfeifer and Kleiter's (2005; 2010) mental probability logic, are empirically adequate computational level theories (Marr, 1982) of reasoning. Specifically, we were interested in whether or not individuals' responses are consistent with the norms proposed by the two accounts. Unfortunately, not all of these results can be used as evidence in favor of these accounts. For p-validity it seems that most of the relevant responses (i.e., responses to MP and MT inferences, as p-validity does not restrict responses to AC and DA) are in fact given in accordance to the norm, for MP all responses were even norm conforming. However, when taking the probability into account that responses could be p-valid by chance by considering the smallest response value that would still be p-valid, the analysis shows that only for MP there is above chance performance. For MT, in contrast, performance was at chance level. Similar results were obtained for the intervals predicted by mental probability logic. When taking the size of the interval as chance level into account, only for MP and, to a lesser degree, for DA did participants responses follow the norm. In contrast, for MT and AC only chance performance was observed. Therefore taken together, only the results of MP and of DA for the coherence based approach can be viewed as evidence for the empirical adequacy of p-validity and mental probability logic.

The probabilized conditional reasoning task, albeit allowing us to run a simultaneous by-subject and by-item analysis on directly obtained estimates of all relevant probabilities, contains features which may have undesirable consequences. For example, the eight questions for each conditional are administered in

one block, which may have led to anchoring or carry-over effects. Additionally, the questions for P(If p then q), P(minor premise), and P(conclusion) were always administered in this order and all the other probabilities afterward which may have exacerbated the above problem or induced order effects (this was one reason for the free parameters in the BL* model). Future research could try to rule our these concerns by for example alternating the order or distributing the items per conditional across the experimental session. Note that the sequence of items in the present experiment was in part necessitated by the requirement to present the probability estimates of the premises in the item asking for the probability of the conclusion. Further, it was the sequence least likely to cause undesirable transfer effects in the probabilized conditional inferences which were of central interest here.

Some new paradigm researchers have argued that, by taking a Bayesian approach in the psychology of reasoning, we will find quite a high level of rationality in people, as judged by Bayesian standards (Oaksford and Chater, 2007, 2009). However, other supporters of the new paradigm are doubtful that the new approach will find a very high degree of rationality in people (Evans and Over, 1996, 2013, and see Elqayam and Evans, 2011). Studies in judgment and decision making have found numerous fallacies and biases in people's probability, and also utility, judgments. In our view, it is excessively optimistic to expect these irrational tendencies to disappear completely when people are using their probability judgments, as they commonly do, in their reasoning. From a dual process perspective, one could predict that there will be an increased tendency for higher level processes to be employed in explicit inferences. These higher processes could increase conformity to normative rules, but do not always, or necessarily, do so, whether the rules are probabilistic or not (Elqayam and Over, 2012; Evans and Stanovich, 2013). We would predict some increase in this conformity, but only expect people to be modestly in line with p-validity and coherence in their reasoning (an expectation also confirmed by Evans et al., 2014).

What we found in our experiment is that people were above chance performance only for the MP inference. This finding needs careful assessment and further study. It appears that people are indeed limited to some extent in how far they conform to Bayesian standards. However, we would point out that MP occupies an absolutely central place in Bayesian inference. Take the classic example of Bayesian inference in a scientific procedure. We infer using Bayes' theorem that there is a conditional probability that a certain hypothesis h holds given evidence e. Recent research, cited above, has shown that people judge the probability of a conditional, P(if e then h), to be the conditional probability, $P(h|e)$. Now the final stage of Bayesian inference is for e to be found true or at least probable to some reasonable degree, so that $P(e)$ is high enough for some confidence in h, $P(h)$, to be inferred. The inference at this last step is usually called conditionalization when $P(h|e)$ is the major premise. We can see that it is an instance of MP when the major premise is the conditional with a degree of belief, P(if e then h) $= P(h|e)$. Bayesian confirmation and belief updating, or belief revision, depend on uses of MP of this general form, when P(if e then h) $= P(h|e)$ is invariant, or rigid, and $P(e)$ is found to be high (see Chater and Oaksford, 2009; Oaksford and Chater, 2013). For this reason, it is significant that we have found MP performance to be above chance level.

ACKNOWLEDGMENTS

This work was supported by Grant KL 614/33-1 to Karl Christoph Klauer and Sieghard Beller from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). We thank Jonathan Evans, Gernot Kleiter, and Niki Pfeifer for many discussions about probabilistic validity and coherence, Jonathan Evans for his analysis that controls for chance responses (Evans et al., 2013), and Igor Douven for comments on this manuscript. Furthermore, we thank David Kellen, Josh O'Brien, and Ronald Reinicke for their help and discussion on **Figure 3**. The article processing charge was funded by the German Research Foundation (DFG) and the Albert Ludwigs University Freiburg in the funding programme Open Access Publishing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00316/abstract

REFERENCES

Adams, E. W. (1998). *A Primer of Probability Logic*. Stanford, CA: Center for the Study of Language and Information.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bull. Psychon. Soc.* 15, 147–149. doi: 10.3758/BF03334492

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Front. Psychol.* 4:328. doi: 10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). *lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version* 1.1-0. Available online at: http://lme4.r-forge.r-project.org/

Bretz, F., Hothorn, T., and Westfall, P. H. (2011). *Multiple Comparisons Using R*. Boca Raton, FL: CRC Press.

Chater, N., and Oaksford, M. (2009). Local and global inferential relations: response to over (2009). *Think. Reason.* 15, 439–446. doi: 10.1080/13546780903361765

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Routledge Academic.

Coletti, G., and Scozzafava, R. (2002). *Probabilistic Logic in a Coherent Setting*. Dordrecht: Kluwer Academic Publishers. doi: 10.1007/978-94-010-0474-9

de Finetti, B. D. (1936). The logic of probability (Translation of 1936 original), translated in R. B. Angell, the logic of probability. *Philos. Stud.* 77, 181–190.

de Finetti, B. D. (1937). "Foresight: its logical laws, its subjective sources (translation of 1937 original,") in *Studies in Subjective Probability*, eds H. E. Kyburg and H. E. Smokler (New York, NY: Wiley), 55–118.

Douven, I., and Verbrugge, S. (2010). The adams family. *Cognition* 117, 302–318. doi: 10.1016/j.cognition.2010.08.015

Douven, I., and Verbrugge, S. (2012). Indicatives, concessives, and evidential support. *Think. Reason.* 18, 480–499. doi: 10.1080/13546783.2012.716009

Douven, I., and Verbrugge, S. (2013). The probabilities of conditionals revisited. *Cogn. Sci.* 37, 711–730. doi: 10.1111/cogs.12025

Edgington, D. (1995). On conditionals. *Mind* 104, 235–329. doi: 10.1093/mind/104.414.235

Elqayam, S., and Evans, J. S. B. T. (2011). Subtracting sought from is: Descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Elqayam, S., and Over, D. E. (2012). Probabilities, beliefs, and dual processing: the paradigm shift in the psychology of reasoning. *Mind Soc.* 11, 27–40. doi: 10.1007/s11299-012-0102-4

Elqayam, S., and Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by elqayam, bonnefon, and over. *Think. Reason.* 19, 249–265. doi: 10.1080/13546783.2013.841591

Evans, J. S. (1993). The mental model theory of conditional reasoning: critical appraisal and revision. *Cognition* 48, 1–20. doi: 10.1016/0010-0277(93)90056-2

Evans, J. S. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Think. Reason.* 18, 5–31. doi: 10.1080/13546783.2011.637674

Evans, J. S. B. T., Handley, S. J., Neilens, H., and Over, D. E. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *Q. J. Exp. Psychol.* 63, 892–909. doi: 10.1080/17470210903111821

Evans, J. S. B. T., Handley, S. J., and Over, D. E. (2003). Conditionals and conditional probability. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 321–335. doi: 10.1037/0278-7393.29.2.321

Evans, J. S. B. T., and Over, D. E. (1996). *Rationality and Reasoning*. Hove: Psychology Press.

Evans, J. S. B. T., and Over, D. E. (2004). *If*. Oxford: OUP. doi: 10.1093/acprof:oso/9780198525134.001.0001

Evans, J. S. B. T., and Over, D. E. (2013). Reasoning to and from belief: deduction and induction are still distinct. *Think. Reason.* 19, 267–283. doi: 10.1080/13546783.2012.745450

Evans, J. S. B. T., and Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Evans, J. S. B. T., Thompson, V. A., and Over, D. E. (2014). *Uncertain Deduction and The New Psychology of Conditional Inference*. Plymouth, UK: University of Plymouth.

Fernbach, P. M., and Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1327–1343. doi: 10.1037/a0031851

Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011). How people interpret conditionals: shifts toward the conditional event. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 635–648. doi: 10.1037/a0022329

Gilio, A. (2002). Probabilistic reasoning under coherence in system p. *Ann. Math. Artif. Intell.* 34, 5–34. doi: 10.1023/A:1014422615720

Gilio, A., and Over, D. E. (2012). The psychology of inferring conditionals from disjunctions: a probabilistic study. *J. Math. Psychol.* 56, 118–131. doi: 10.1016/j.jmp.2012.02.006

Halekoh, U., and Højsgaard (2013). *pbkrtest: Parametric Bootstrap and Kenward Roger Based Methods for Mixed Model Comparison*. *R Package Version* 0.3-5.1. Available online at: http://people.math.aau.dk/~sorenh/software/pbkrtest/

Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. (2006). A lego system for conditional inference. *Am. Stat.* 60, 257–263. doi: 10.1198/000313006X118430

Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. (2008). Implementing a class of permutation tests: the coin package. *J. Stat. Softw.* 28, 1–23.

Johnson-Laird, P. N., and Byrne, R. M. (1991). *Deduction*. Hove: Lawrence Erlbaum Associates, Inc.

Johnson-Laird, P. N., and Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychol. Rev.* 109, 646–678. doi: 10.1037//0033-295X.109.4.646

Judd, C. M., Westfall, J., and Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *J. Pers. Soc. Psychol.* 103, 54–69. doi: 10.1037/a0028347

Klauer, K. C., Beller, S., and Hütter, M. (2010). Conditional reasoning in context: a dual-source model of probabilistic inference. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 298–323. doi: 10.1037/a0018705

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philos. Rev.* 85, 297–315. doi: 10.2307/2184045

Liu, I.-m. (2003). Conditional reasoning and conditionalization. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 694–709. doi: 10.1037/0278-7393.29.4.694

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H. Freeman.

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101, 608–631. doi: 10.1037/0033-295X.101.4.608

Oaksford, M., and Chater, N. (2001). The probabilistic approach to human reasoning. *Trends Cogn. Sci.* 5, 349–357. doi: 10.1016/S1364-6613(00)01699-5

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780198524496.001.0001

Oaksford, M., and Chater, N. (2009). Précis of bayesian rationality: the probabilistic approach to human reasoning. *Behav. Brain Sci.* 32, 69–84. doi: 10.1017/S0140525X09000284

Oaksford, M., and Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Think. Reason.* 19, 346–379. doi: 10.1080/13546783.2013.808163

Oaksford, M., Chater, N., and Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 883–899. doi: 10.1037/0278-7393.26.4.883

Oberauer, K., Geiger, S. M., Fischer, K., and Weidenfeld, A. (2007). Two meanings of if? individual differences in the interpretation of conditionals. *Q. J. Exp. Psychol.* 60, 790–819. doi: 10.1080/17470210600822449

Oberauer, K., and Wilhelm, O. (2003). The meaning(s) of conditionals: conditional probabilities, mental models, and personal utilities. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 680–693. doi: 10.1037/0278-7393.29.4.680

Over, D. E. (2009). New paradigm psychology of reasoning. *Think. Reason.* 15, 431–438. doi: 10.1080/13546780903266188

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., and Sloman, S. A. (2007). The probability of causal conditionals. *Cogn. Psychol.* 54, 62–97. doi: 10.1016/j.cogpsych.2006.05.002

Pfeifer, N. (2013). The new psychology of reasoning: a mental probability logical perspective. *Think. Reason.* 19, 329–345. doi: 10.1080/13546783.2013.838189

Pfeifer, N. (2014). Reasoning about uncertain conditionals. *Studia Logica*. Available online at: http://link.springer.com/article/10.1007%2Fs11225-013-9505-4

Pfeifer, N., and Kleiter, G. D. (2005). Towards a mental probability logic. *Psychol. Belgica* 45, 71–99. doi: 10.5334/pb-45-1-71

Pfeifer, N., and Kleiter, G. D. (2006). Inference in conditional probability logic. *Kybernetika* 42, 391–404.

Pfeifer, N., and Kleiter, G. D. (2007). "Human reasoning with imprecise probabilities: modus ponens and denying the antecedent," in *Proceedings of the 5th International Symposium on Imprecise Probability: Theories and Applications* (Prague), 347–356.

Pfeifer, N., and Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *J. Appl. Logic* 7, 206–217. doi: 10.1016/j.jal.2007.11.005

Pfeifer, N., and Kleiter, G. D. (2010). "The conditional in mental probability logic," in *Cognition and Conditionals: Probability and Logic in Human Thought*, eds M. Oaksford N. and Chater (Oxford: Oxford University Press), 153–173.

Politzer, G., Over, D. E., and Baratgin, J. (2010). Betting on conditionals. *Think. Reason.* 16, 172–197. doi: 10.1080/13546783.2010.504581

Ramsey, F. P. (1931). *The Foundations of Mathematics and Other Logical Essays*. New York, NY: K. Paul, Trench, Trubner & Co.; Harcourt, Brace and Co., London.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rips, L. J. (2001). Two kinds of reasoning. *Psychol. Sci.* 12, 129–134. doi: 10.1111/1467-9280.00322

Rottman, B. M., and Hastie, R. (2014). Reasoning about causal relationships: inferences on causal networks. *Psychol. Bull.* 140, 109–139. doi: 10.1037/a0031903

Schroyens, W. J., and Schaeken, W. (2003). A critique of oaksford, chater, and larkin's (2000) conditional probability model of conditional reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 140–149. doi: 10.1037/0278-7393.29.1.140

Singmann, H. (2013). *afex: Analysis of Factorial Experiments*. *R Package Version* 0.6-77/r77. Available online at: http://www.psychologie.uni-freiburg.de/Members/singmann/R/afex

Singmann, H., and Klauer, K. C. (2011). Deductive and inductive conditional inferences: two modes of reasoning. *Think. Reason.* 17, 247–281. doi: 10.1080/13546783.2011.572718

Sloman, S. A. (2005). *Causal Models: How People Think About The World and Its Alternatives*. Oxford, NY: Oxford University Press.

Stevenson, R. J., and Over, D. E. (2001). Reasoning from uncertain premises: effects of expertise and conversational context. *Think. Reason.* 7, 367–390. doi: 10.1080/13546780143000080

Tversky, A., and Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* 5, 297–323. doi: 10.1007/BF00122574

Wagner, C. G. (2004). Modus tollens probabilized. *Br. J. Philos. Sci.* 55, 747–753. doi: 10.1093/bjps/55.4.747

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Stat. Med.* 22, 3527–3541. doi: 10.1002/sim.1572

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 November 2013; paper pending published: 11 February 2014; accepted: 26 March 2014; published online: 17 April 2014.

*Citation: Singmann H, Klauer KC and Over D (2014) New normative standards of conditional reasoning and the dual-source model. Front. Psychol. **5**:316. doi: 10.3389/fpsyg.2014.00316*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

OPINION ARTICLE
published: 12 March 2014
doi: 10.3389/fpsyg.2014.00217

Modeling causal conditional reasoning data using SDT: caveats and new insights

Dries Trippas*, Michael F. Verde, Simon J. Handley, Matthew E. Roser, Nicolas A. McNair and Jonathan St. B. T. Evans

Faculty of Health and Human Sciences, School of Psychology, Cognition Institute, Plymouth University, Plymouth, UK
**Correspondence: dries.trippas@gmail.com*

Edited by:
David E. Over, Durham University, UK

Reviewed by:
Henrik Singmann, Albert-Ludwigs-Universität Freiburg, Germany
Evan Heit, University of California, Merced, USA

Keywords: causal conditionals, reasoning, signal detection theory, belief bias, normative models

In deductive reasoning, people are asked to infer the truth of an argument's conclusion given a set of premises. Research into the processes underlying deduction has focused on examining how well people discriminate between logically valid and invalid arguments, and how irrelevant factors such as one's prior beliefs interfere with the ability to reason logically (Evans et al., 1983). This normative approach to validity has traditionally informed both practice and theory in the literature. However, its critics argue that "normativism" often leads investigators to biased or misleading interpretations of phenomena (Elqayam and Evans, 2011).

Formal modeling of deductive reasoning has often been successful by taking the traditional, normative approach. A case in point is the application of signal detection theory (SDT; Macmillan and Creelman, 2005) to the investigation of belief bias in syllogistic reasoning (Dube et al., 2010). In the SDT model, deductive judgments are based on strength of evidence; an argument is judged to be valid if its strength exceeds a criterion value. Because the choice of criterion is independent of the ability to discriminate between classes of arguments, the SDT model makes it possible to isolate response bias from accuracy. Dube et al. examined these two factors using ROC curves, which plot hits against false alarms at several levels of confidence. Hits and false alarms were defined in normative fashion as responding "valid" to logically valid and logically invalid conclusions, respectively.

Their analysis of ROCs led them to argue two significant points. First, contrary to prevailing theories of belief bias, conclusion believability can affect response bias without affecting the quality of reasoning. Second, the curvilinear shape of the ROCs is consistent with the distributional assumptions of SDT. The latter is a key test because finding linear rather than curvilinear ROCs would be problematic for the model. The curvilinear ROCs obtained in syllogistic (see also Dube et al., 2011; Trippas et al., 2013; but see Klauer and Kellen, 2011) and other forms of reasoning (Heit and Rotello, 2010, 2014) are similar to those widely observed in memory and perception (Pazzaglia et al., 2013). This consistency across domains strengthens the case for the usefulness of the SDT approach. It also leads to an expectation of similar findings in other areas of reasoning. Below, we describe findings from conditional reasoning that violate this expectation in a surprising yet enlightening way.

Causal conditionals are a form of deduction prevalent in everyday life. Consider the proposition: "If healthy foods are cheaper, then more people will eat healthy foods." Four types of conditional inferences are possible: modus ponens (MP; "Healthy foods are cheaper, therefore more people will eat healthy foods"), modus tollens (MT; "Fewer people eat healthy foods, therefore healthy foods are not cheaper"), affirmation of the consequent (AC; "More people eat healthy foods, therefore healthy foods are cheaper"), and denial of the antecedent (DA; "Healthy foods are not cheaper, therefore less people eat healthy foods").

From a normative point of view, MP and MT are valid and AC and DA are invalid inferences. Theories differ as to how people determine validity in these problems. According to mental model theory (Johnson-Laird and Byrne, 2002), people construct an initial mental model of the conditional (e.g., p q) which may then be fleshed out by considering additional models (not-p q; not-p not-q). According to the suppositional account of the conditional (Evans et al., 2003, 2005; Evans and Over, 2004, 2012), people evaluate the subjective probability of a conditional by hypothetically supposing p and then assessing the conditional probability of q given p, P(q|p). This relation between the natural language conditional and the conditional probability, P(if p then q) = P(q|p), can be used in a Bayesian/probabilistic model of conditional inference (Oaksford et al., 2000; Oaksford and Chater, 2009, 2013).

What these theories have in common is that there is no fundamental difference in how people process affirmation (MP + AC) and denial (MT + DA) inferences. This makes an SDT analysis straightforward and no different to that taken with the study of belief bias in syllogistic reasoning. For our case study, we analyzed aspects of a data set collected as part of a larger project under the direction of the fourth author of this paper[1]. This study examined the influence of belief in causal conditional problems (e.g., believable: "If oil prices continue to rise, then UK

[1]This research was supported by the award of an ESRC project grant RES-062-23-3285.

petrol prices will rise"; unbelievable: "If global temperatures rise, then less arctic ice will melt"). Hits were defined as "valid" responses to MP and MT and false alarms were defined as "valid" responses to AC and DA. This produced the ROCs seen in the top panel of **Figure 1**. The results are similar in some respects to those reported by Dube et al. (2010) for syllogisms: believability had no effect on accuracy (ROCs for believable and unbelievable items fall on the same curve) but seemed to affect response bias (confidence criteria for believable items are shifted to the right)[2]. However, there is a surprising difference: in contrast to the curvilinear ROCs observed with syllogisms, conditionals produced linear ROCs. A linear regression of the ROC (collapsing over believability) provided a good fit, $R^2 = 99.9\%$. Adding a quadratic component did not improve the fit, $p = 0.78$. Taken at face value, this result suggests that conditional reasoning requires a profoundly different model than the one that has seemed so successful when applied to other forms of reasoning, not to mention other cognitive tasks.

A different picture emerges when we depart from the strictly normative approach and consider separately how people respond to affirmation and denial conditionals. In the bottom left panel of **Figure 1**, plotting MP (hits) against AC (false alarms) yields typically curvilinear ROCs. Linear regression (collapsing over believability) provided a fit, $R^2 = 96\%$, that was significantly improved by the

[2]Note that the current data pattern does not necessitate a criterion shift, but that it is also consistent with a symmetrical distribution shift. For more discussion on this issue, see, e.g., Verde et al. (2010).

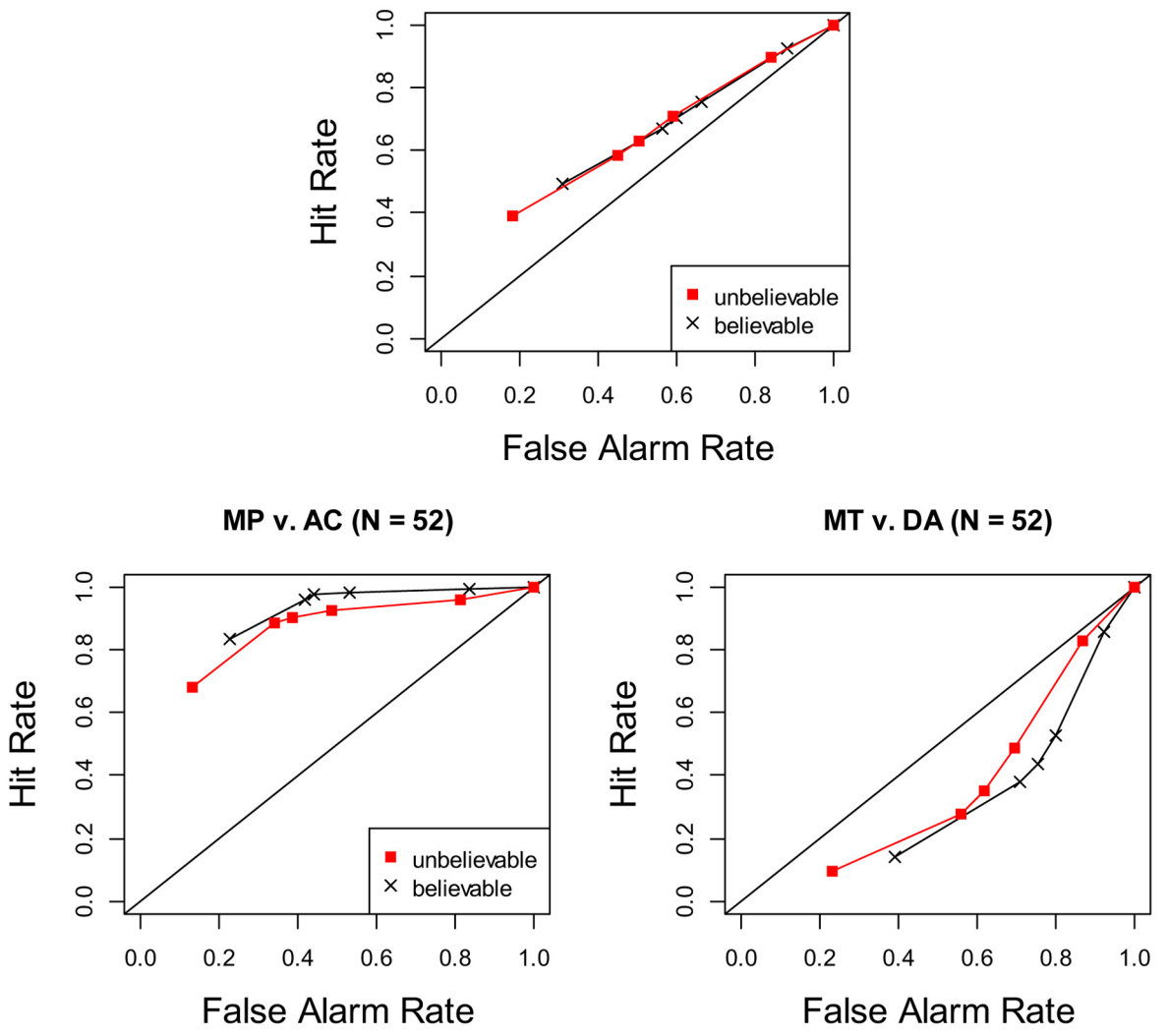


FIGURE 1 | ROC curves of causal conditionals. Top panel: Valid (MP + MT) vs. invalid (AC + DA). **Bottom left:** affirmation conditionals (MP vs. AC). **Bottom right:** denial conditionals (MT vs. DA). Points on the ROC imply a more liberal response criterion (lower confidence responses) for identical levels of sensitivity. The points are plotted cumulatively such that the leftmost point = high confidence hits vs. false alarms, with the next point down being high + medium confidence hits vs. false alarms, and so forth.

addition of a quadratic component, $R^2 = 99.99\%$, $p < 0.004$. Accuracy is defined by the distance of the ROCs from the chance diagonal. Contrary to the poor accuracy on display in the aggregate results in the top panel, people are quite sensitive to argument structure when affirmation is involved. In the bottom right panel of **Figure 1**, plotting MT (hits) against DA (false alarms) again yields typically curvilinear ROCs. Linear regression (collapsing over believability) provided a fit, $R^2 = 98\%$, that was significantly improved by the additional of a quadratic component, $R^2 = 99.99\%$, $p < 0.002$. People were sensitive to argument structure, but the position of the ROCs below the diagonal indicates that their treatment of denial arguments departed from the normative; MT are treated as *less* valid than AC.

Applying the SDT model in a normative fashion, as would seem reasonable given extant theories of conditional reasoning, produced results that contrast sharply with previous findings. The clearly linear ROC in the top panel of **Figure 1** is not only unlike the curvilinear ROCs observed with syllogisms but if taken at face value is problematic for the SDT model. It could be that there is something fundamentally different in the way that people reason about causal conditionals as compared to other types of problems. It seems to us more likely that the difference lies with affirmation and denial inferences; the latter do not seem to be treated in the normatively prescribed fashion. Once this is assumed, the ROC results become more sensible and fall in line with previous results (in a reanalysis of published and unpublished data sets, Heit and Rotello, 2014, have also reported curvilinear ROCs from MP plotted in the manner of **Figure 1**, lower left). This interpretation converges with Singmann and Klauer's (2011) finding, based on state-trace analysis, that affirmation and denial problems may depend on different processes.

Why use ROC analysis rather than simply examine the raw validity judgments? Interpreting the latter often relies on assumptions that may not be justified (Klauer et al., 2000; Dube et al., 2010). The main advantage of a formal model like SDT lies in its specification of assumptions. However, models can also produce insights that are not obvious at first glance. A qualitative difference between affirmation and denial inferences is not necessarily predicted by extant theories. Moreover, various manipulations seem to exert a similar effect on both types of inferences (e.g., Cummins, 1995). Finally, it is interesting to note that the production of linear ROCs when performance is driven by multiple underlying processes has been predicted in theory (DeCarlo, 2002). These results may offer a case study of how this can occur in practice.

REFERENCES

Cummins, D. D. (1995). Naive theories and causal deduction. *Mem. Cognit.* 23, 646–658. doi: 10.3758/BF03197265

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: theoretical developments with applications to recognition memory. *Psychol. Rev.* 109, 710–721. doi: 10.1037/0033-295X.109.4.710

Dube, C., Rotello, C. M., and Heit, E. (2010). Assessing the belief bias effect with ROCs: it's a response bias effect. *Psychol. Rev.* 117, 831–863. doi: 10.1037/a0019634

Dube, C., Rotello, C. M., and Heit, E. (2011). The belief bias effect is aptly named: a reply to Klauer and Kellen (2010). *Psychol. Rev.* 118, 155–163. doi: 10.1037/a0021774

Elqayam, S., and Evans, St. B. T. (2011). Subtracting 'ought' from 'is': descriptivism versusnormativism in the study of the human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Evans, J. St. B. T., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Mem. Cognit.* 11, 295–306. doi: 10.3758/BF03196976

Evans, J. St. B. T., Handley, S. J., and Over, D. E. (2003). Conditionals and conditional probability. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 321–355. doi: 10.1037/0278-7393.29.2.321

Evans, J. St. B. T., and Over, D. E. (2004). *If.* Oxford: Oxford University Press.

Evans, J. St. B. T., and Over, D. E. (2012). Reasoning to and from belief: deduction and induction are still distinct. *Think. Reas.* 19, 267–283. doi: 10.1080/13546783.2012.745450

Evans, J. St. B. T., Over, D. E., and Handley, S. J. (2005). Suppositions, extensionality, and conditionals: a critique of the mental model theory of Johnson-Laird and Byrne (2002). *Psychol. Rev.* 112, 1040–1052. doi: 10.1037/0033-295X.112.4.1040

Heit, E., and Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 805–812. doi: 10.1037/a0018784

Heit, E., and Rotello, C. M. (2014). Traditional difference-score analyses of reasoning are flawed. *Cognition* 131, 75–91. doi: 10.1016/j.cognition.2013.12.003

Johnson-Laird, P. N., and Byrne, R. M. J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychol. Rev.* 109, 646–678. doi: 10.1037/0033-295X.109.4.646

Klauer, K. C., and Kellen, D. (2011). Assessing the belief bias effect with ROCs: reply to Dube, Rotello, and Heit (2010). *Psychol. Rev.* 118, 155–164. doi: 10.1037/a0020698

Klauer, K. C., Musch, J., and Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychol. Rev.* 107, 852–884. doi: 10.1037/0033- 295X.107.4.852

Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide, 2nd Edn.* Mahwah, NJ: Erlbaum.

Oaksford, M., and Chater, N. (2009). Précis of Bayesian rationality: the probabilistic approach to human reasoning. *Behav. Brain Sci.* 32, 69–84. discussion: 85–120. doi: 10.1017/S0140525X09000284

Oaksford, M., and Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Think. Reason.* 19, 346–379. doi: 10.1080/13546783.2013.808163

Oaksford, M., Chater, N., and Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 883–889. doi: 10.1037/0278-7393.26.4.883

Pazzaglia, A. M., Dube, C., and Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: implications for recognition and beyond. *Psychol. Bull.* 139, 1173–1203. doi: 10.1037/a0033044

Singmann, H., and Klauer, K. C. (2011). Deductive and inductive conditional inferences: two modes of reasoning. *Think. Reason.* 17, 247–281. doi: 10.1080/13546783.2011.572718

Trippas, D., Handley, S. J., and Verde, M. F. (2013). The SDT model of belief bias: complexity, time, and cognitive ability mediate the effects of believability. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1393–1402. doi: 10.1037/a0032398

Verde, M. F., Stone, L. K., Hatch, H. S., and Schnall, S. (2010). Distinguishing between mnemonic and attributional sources of familiarity: positive emotion bias as a case study. *Mem. Cognit.* 38, 142–153. doi: 10.3758/MC.38.2.142

Received: 30 November 2013; accepted: 25 February 2014; published online: 12 March 2014.

Citation: Trippas D, Verde MF, Handley SJ, Roser ME, McNair NA and Evans JSBT (2014) Modeling causal conditional reasoning data using SDT: caveats and new insights. Front. Psychol. 5:217. doi: 10.3389/fpsyg.2014.00217

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

GENERAL COMMENTARY
published: 14 May 2014
doi: 10.3389/fpsyg.2014.00402

Concerns with the SDT approach to causal conditional reasoning: a comment on Trippas, Handley, Verde, Roser, McNair, and Evans (2014)

Henrik Singmann*† and David Kellen*†

Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany
**Correspondence: henrik.singmann@psychologie.uni-freiburg.de; david.kellen@psychologie.uni-freiburg.de*

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Dries Trippas, Plymouth University, UK
Richard Donald Morey, University of Groningen, Netherlands

† These authors have contributed equally to this work.

Keywords: conditional reasoning, syllogistic reasoning, belief bias, signal detection models, measurement models, model identifiability

A commentary on

Modeling causal conditional reasoning data using SDT: caveats and new insights
by Trippas, D., Verde, M. F., Handley, S. J., Roser, M. E., McNair, N. A., and Evans, J. S. B. T. (2014). Front. Psychol. 5:217. doi: 10.3389/fpsyg.2014.00217

Signal Detection Theory (SDT; Wickens, 2002) is a prominent measurement model that characterizes observed classification responses in terms of discriminability and response bias. In recent years, SDT has been increasingly applied within the psychology of reasoning (Rotello and Heit, 2009; Dube et al., 2010; Heit and Rotello, 2010, 2014; Trippas et al., 2013). SDT assumes that different stimulus types (e.g., valid and invalid syllogisms) are associated with different (presumably Gaussian) evidence or argument-strength distributions. Responses (e.g., "Valid" and "Invalid") are produced by comparing the argument-strength of each syllogism with a set of established response criteria (**Figure 1A**). The response profile associated to each stimulus type can be represented as a Receiver Operating Charateristics (ROC) function by plotting performance pairs (i.e., hits and false-alarms) along different response criteria, which Gaussian SDT predicts to be curvilinear (**Figure 1B**).

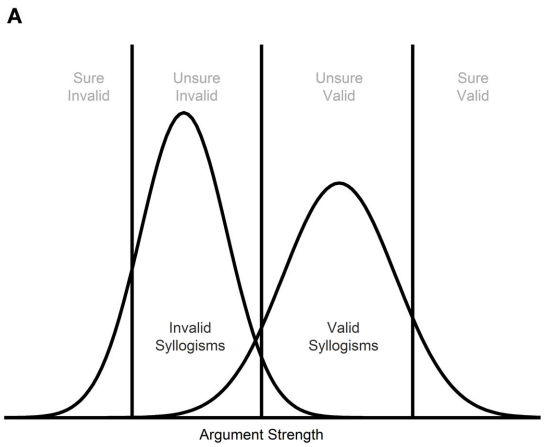
Trippas et al. (2014; henceforth THVRME) applied SDT to causal-conditional reasoning and make two points: (1) that SDT provides an informative characterization of data from a reasoning experiment with two orthogonal factors such as believability and argument validity; (2) that an inspection of the shape of causal-conditional ROCs provides insights on the suitability of normative theories with the consequence to consider affirmation and denial problems separately.

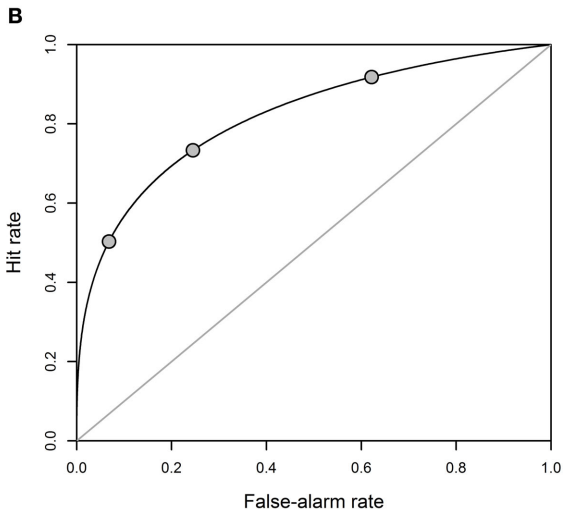
The goal of this comment is to make two counterarguments: First, to point out that the SDT model is often unable to provide an informative characterization of data in designs as discussed by THVRME as it fails to unambiguously separate argument strength and response bias. THVRME's conclusion that "believability had no effect on accuracy [. . .] but seemed to affect response bias" (p. 4) solely hinge on arbitrary assumptions. Second, that THVRME's reliance on ROC shape to justify a separation between affirmation and denial problems is unnecessary and misguided.

1. SEPARATING ARGUMENT STRENGTH AND RESPONSE BIAS

Assume a toy SDT model with four (equal-variance) evidence distributions, corresponding to the four types of syllogisms resulting from the Validity (V = Valid/I = Invalid) × Believability (B = Believable/U = Unbelievable) factorial design. Now, let the means of the distributions be given by the main effects of Validity and Believability as well as their interaction, using a 0/1 factor coding. This factorial design produces the table in **Figure 1C**.

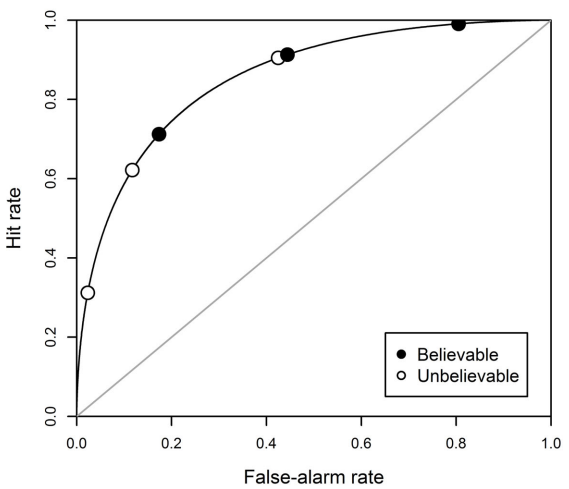
The possibility of specifying different response criteria for the two levels of the Believability factor leads to an unidentifiable SDT model in which differences between means trade-off with differences between response criteria (Wickens and Hirshman, 2000; Klauer and Kellen, 2011). For example, the ROCs in **Figure 1D** can be equally accounted for by a difference in the distributions (**Figure 1E**) or by a response-criteria shift (**Figure 1F**). Because THVRME and others fix IB to 0 a priori, they *enforce* a response-criteria shift interpretation of the ROCs. This ambiguity in the characterization of the data compromises the attempt to relate its parameters with different accounts on e.g., the belief-bias effect. THVRME briefly mention this (see their Footnote 2) but do not address its implications. The IB = 0 restriction implies that effects of believability on argument strength can *only* be detected if the interaction term is non-zero as the main-effect term of believability is effectively censored. This means that a pure criteria-shift account can be enforced as long as no severe violations of additivity (i.e., an interaction) are observed. In other words, only when VB differs from VU (while assuming IB = 0) can the proposed pure criteria-shift model be rejected. To make matter worse, the criteria-shift account is implausible to begin with given that it runs counter to empirical work showing that individuals do not tend to change their response criteria on a trial-by-trial basis (e.g., Morrell et al., 2002).

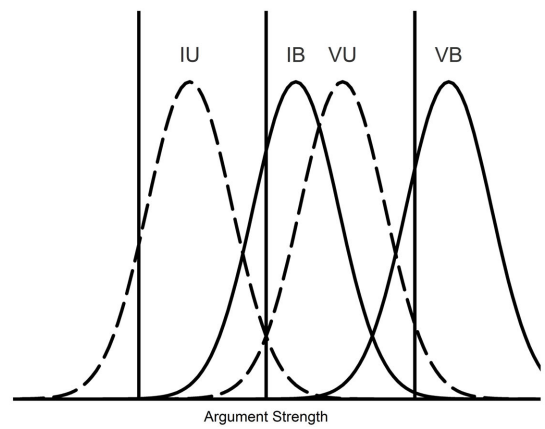




C

	Unbelievable (**U**)	Believable (**B**)
Invalid (**I**)	**IU** 0	**IB** Main-effect Believability
Valid (**V**)	**VU** Main-effect Validity	**VB** Main-effect Validity + Main-effect Believability + Interaction





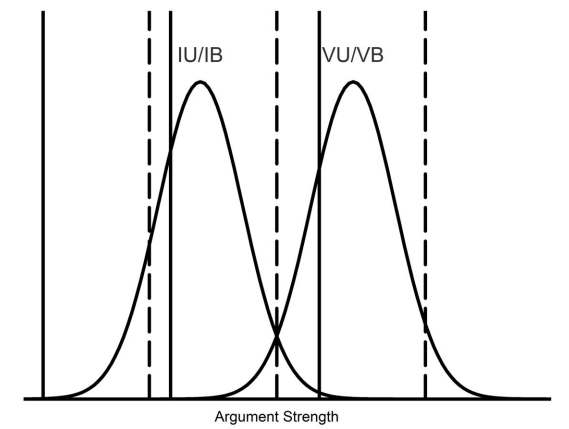


FIGURE 1 | (A) A graphical representation of the SDT model for a syllogistic reasoning task. **(B)** ROC curve representing the cumulative probabilities for hypothetical pairs of hits and false-alarms ("valid" responses to valid and invalid syllogisms, respectively) based on the four response categories depicted in **(A)**. **(C)** Factorial design of Believability × Validity representing the means of the SDT evidence distributions. **(D)** ROCs for believable and unbelievable syllogisms. **(E)** Distribution shift account of ROCs in which the distributions for believable syllogisms (solid lines) are shifted to the right. **(F)** Response-criteria shift account of ROCs in which the response criteria for believable syllogisms (solid lines) are shifted to the left. Note that for ease in the illustration the response proportions implied by the SDT accounts of panels **(E,F)** do not exactly correspond to the response proportions depicted in panel **(D)**.

2. DATA AGGREGATION CONFOUNDS IN CAUSAL-CONDITIONAL REASONING

THVRME's reliance on ROC shape to justify the separation between the affirmation and denial problems is *unnecessary* and *misguided*: It is unnecessary because the acceptance rates (A) already show the pattern $A_{MP} > A_{DA}$ and $A_{AC} < A_{MT}$[1], indicating that performance is "above chance" for affirmation problems but "below chance" for denial problems (see Singmann and Klauer, 2011, for similar results). This contrasting pattern in the acceptance rates alone indicates that aggregating affirmation and denial problems is an unwise option. Note that the criticisms associated to acceptance rates (e.g., Klauer et al., 2000; Dube et al., 2010; Heit and Rotello, 2014) do not hold here as they are exclusively concerned with the interpretation of response patterns of the form $A_{VB} > A_{VU}, A_{IB} > A_{IU}$.

THVRME's use of eyeball and regression-based evaluations of ROC shape is misguided because it overlooks the more subtle (but still pernicious) distortions from item heterogeneity (Rouder and Lu, 2005), but also because it fails to characterize SDT's *actual* ability to fit their own data. As it turns out, SDT fits the linear aggregate ROCs better (VB/IB: $G^2(3) = 7.95$, $p = 0.05$; VU/IU: $G^2(3) = 10.63$, $p = 0.01$) than the curvilinear ROCs from acceptance and denial problems (smallest $G^2(3) = 13.51$, $p < 0.01$). The sufferable fit of the aggregate data is not surprising given Gaussian SDT's ability to account for near-linear ROCs when performance is low[2].

3. CONCLUSION

THVRME attempt to demonstrate the value of SDT modeling in research on causal-conditional reasoning. However, the main motivation for employing SDT is to characterize differences in argument-strength and response bias across conditions. As we have shown, the approach of THVRME is unable to accomplish this in an unambiguous fashion. Furthermore, THVRME's detection of differences between affirmation and denial problems hinges on an evaluation of ROC shape that is not only unnecessary (as acceptance rates are sufficient) but also fails to relate ROCs with SDT predictions in a principled way. SDT has a long and successful history in psychological research, and will likely provide important insights in the reasoning domain; however, from the current standpoint, we fail to see the exact contribution of the SDT modeling advocated by THVRME and others (e.g., Dube et al., 2010; Trippas et al., 2013; Heit and Rotello, 2014) to research on human reasoning.

[1] MP, Modus Ponens; MT, Modus Tollens; AC, Afirmation of the Consequent; DA, Denial of the Antecendent.

[2] Note that a non-parametric characterization of ROCs is possible (Kornbrot, 2006).

ACKNOWLEDGMENTS

We thank Matt Roser, Nicolas McNair, and Jonathan Evans for providing their aggregated data. This work was supported by Grant KL 614/33-1 to Karl Christoph Klauer and Sieghard Beller from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

REFERENCES

Dube, C., Rotello, C. M., and Heit, E. (2010). Assessing the belief bias effect with ROCs: it's a response bias effect. *Psychol. Rev.* 117, 831–863. doi: 10.1037/a0019634

Heit, E., and Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 805–812. doi: 10.1037/a0018784

Heit, E., and Rotello, C. M. (2014). Traditional difference-score analyses of reasoning are flawed. *Cognition* 131, 75–91. doi: 10.1016/j.cognition.2013.12.003

Klauer, K. C., and Kellen, D. (2011). Assessing the belief bias effect with ROCs: reply to dube, rotello, and heit (2010). *Psychol. Rev.* 118, 164–173. doi: 10.1037/a0020698

Klauer, K. C., Musch, J., and Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychol. Rev.* 107, 852–884. doi: 10.1037/0033-295X.107.4.852

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: model-based and distribution-free measures and evaluation. *Percept. Psychophys.* 68, 393–414. doi: 10.3758/BF03193685

Morrell, H. E. R., Gaitan, S., and Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 1095–1110. doi: 10.1037/0278-7393.28.6.1095

Rotello, C. M., and Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1317–1330. doi: 10.1037/a0016648

Rouder, J. N., and Lu, J. (2005). An introduction to bayesian hierarchical models with an application in the theory of signal detection. *Psychon. Bull. Rev.* 12, 573–604. doi: 10.3758/BF03196750

Singmann, H., and Klauer, K. C. (2011). Deductive and inductive conditional inferences: two modes of reasoning. *Think. Reason.* 17, 247–281. doi: 10.1080/13546783.2011.572718

Trippas, D., Handley, S. J., and Verde, M. F. (2013). The SDT model of belief bias: complexity, time, and cognitive ability mediate the effects of believability. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1393–1402. doi: 10.1037/a0032398

Trippas, D., Handley, S. J., Verde, M. F., Roser, M., Mcnair, N., and Evans, J. S. B. T. (2014). Modeling causal conditional reasoning data using SDT: caveats and new insights. *Front. Psychol.* 5:217. doi: 10.3389/fpsyg.2014.00217

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford, NY: Oxford University Press.

Wickens, T. D., and Hirshman, E. (2000). False memories and statistical design theory: comment on miller and wolford (1999) and roediger and McDermott (1999). *Psychol. Rev.* 107, 377–383. doi: 10.1037/0033-295X.107.2.377

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 March 2014; paper pending published: 26 March 2014; accepted: 16 April 2014; published online: 14 May 2014.

*Citation: Singmann H and Kellen D (2014) Concerns with the SDT approach to causal conditional reasoning: a comment on Trippas, Handley, Verde, Roser, McNair, and Evans (2014). Front. Psychol. **5**:402. doi: 10.3389/fpsyg.2014.00402*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

GENERAL COMMENTARY
published: 19 February 2015
doi: 10.3389/fpsyg.2015.00184

Alleviating the concerns with the SDT approach to reasoning: reply to Singmann and Kellen (2014)

Dries Trippas[1]*, Michael F. Verde[2] and Simon J. Handley[2]

[1] *Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany*
[2] *School of Psychology, Cognition Institute, Plymouth University, Plymouth, UK*
**Correspondence: trippas@mpib-berlin.mpg.de*

Edited by:
David E. Over, Durham University, UK

Reviewed by:
Shira Elqayam, De Montfort University, UK

Keywords: reasoning, signal detection theory, model fitting, model identifiability, belief bias, response bias

A commentary on

Concerns with the SDT approach to causal conditional reasoning: a comment on Trippas, Verde, Handley, Roser, McNair, and Evans (2014).
by Singmann, H., and Kellen, D. (2014). Front. Psychol. 5:402. doi: 10.3389/fpsyg.2014.00402

In their comment on our article (Trippas et al., 2014a), Singmann and Kellen (2014; henceforth SK) suggest that our use of signal detection theory (SDT) provides an uninformative characterization of the data, that our application of SDT to causal-conditional reasoning is unnecessary and misguided, and that the model does not provide a good fit of the data. We will address each of these points.

SK's concern that our use of SDT is uninformative rests on a single issue: how to interpret the shift in the location of the confidence points along the ROC when comparing the believable and unbelievable conditions (Trippas et al., 2014a; Figure 1), a shift that in real terms represents a greater tendency to accept believable arguments as "valid." We find SK's focus on this aspect of the data surprising because it has no direct bearing on the main points of the article, which have to do with the changes in the shape and separation of the ROCs when we segregate the data in different ways (Trippas et al., 2014a, Figure 1, comparing the top and bottom panels). For this reason, we mentioned the confidence point shift only once, saying that it fits a pattern previously interpreted by Dube et al. (2010) as a shift in response bias, but which might also be due to a symmetric shift in the evidence distributions. This is a succinct way of stating what SK describe in great detail in their toy model. We have no problem with the fact that the confidence point shift has alternative interpretations. Although it is not integral to the thrust of the article, this aspect of the data is worth noting because the pattern is observed in other reasoning tasks and represents a point of continuity despite the apparent discontinuity in other aspects of the ROC data.

We gather that SK's focus on the unidentifiability issue is meant to be a critique of the SDT model in general. In our view, it is not a compelling critique. Following convention, we use "accuracy" to denote sensitivity, the ability to discriminate classes of items (valid from invalid). Accuracy depends on the relative distance between the valid and invalid distributions. If some factor were to increase the argument strength of invalid and valid arguments by exactly the same degree, accuracy would remain constant. This is a specific circumstance which the SDT model cannot distinguish from a shift in response bias. The model is, however, unambiguous in distinguishing between changes in accuracy from those that might be ascribed solely to response bias. This is where the theoretical power of the model lies (e.g., Trippas et al., 2013).

As for the question of response bias, the SDT model is widely used in domains like memory where criterion placement is an issue because theorists have a range of other tools at their disposal to deal with ambiguity (they can, for example, use manipulations that plausibly only affect response bias). In their final point, SK cite work in recognition memory (Morrell et al., 2002) to argue that trial-by-trial criterion shifts are implausible. These findings describe the specific case in which test stimuli are indistinguishable save for an internal signal of mnemonic strength. When the stimuli are overtly distinguishable on other dimensions, people seem quite capable of shifting their criterion from one trial to the next (Dobbins and Kroll, 2005; Aminoff et al., 2012). Whether people do use different response criteria when judging believable and unbelievable arguments remains an open question, but the memory literature provides ample reason to believe that it is plausible.

SK argue against the application of SDT to conditional reasoning because one can reach the same conclusions by examining raw acceptance rates. This misses the point of using a model like SDT, which is to view the data within a consistent, theoretically justified framework. The problems that can arise when raw acceptance rates are used to measure accuracy are well documented (Klauer et al., 2000; Dube et al., 2010; Heit and Rotello, 2014) and certainly apply here.

SK make a good point in observing that the fit of the model to Roser and colleagues' ROC curves is poor. The problem may lie in the application of the model to aggregated data. It is well known that G^2 depends on sample size such that aggregate model fits very often lead to violations of absolute fit. One alternative approach is to evaluate model fit for each participant individually (Cohen et al., 2008).

To demonstrate the problematic nature of assessing the model fit of aggregated samples in terms of G^2 when sample sizes are large, we combined data from 131 participants from previously published work on belief bias (Trippas et al., 2013, 2014b). We fit the believable and unbelievable ROCs separately, both aggregated and on a per-participant basis. The aggregate fit to the believable ROC was borderline acceptable, $G^2_{(3)} = 6.97$, $p = 0.07$. For the unbelievable ROC, the fit was unacceptable, $G^2_{(3)} = 50.6$, $p < 0.001$. The individual fits paint a prettier picture: for the believable problems, only 9 out of 131 or less than 7% of the participants show a violation of fit ($p < 0.05$). The unbelievable-problems case fares even better, with only 4 out of 131 or about 3% of the participants producing ill-fitting data patterns. How can such drastically different patterns of fit emerge? Individual differences potentially play a large role: unbelievable problems elicit different reasoning strategies in different people (Trippas et al., 2013, 2014b,c), and aggregating across such data patterns will suggest that the model is inappropriate. We suspect that similar factors have contributed to the poor fits reported by SK.

SK's comments speak to a number of interesting issues that deserve to be raised in the wider discussion surrounding the SDT approach to modeling human reasoning. It is useful, however, to reiterate the point of our original article which seems to be lost in the discussion of side issues. A strict adherence to "normativism" often leads investigators to biased or misleading interpretations of phenomena (Elqayam and Evans, 2011). The default normative approach to the application of SDT to reasoning illustrates precisely this problem in the case of causal conditionals.

REFERENCES

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., et al. (2012). Individual differences in shifting decision criterion: a recognition memory study. *Mem. Cogn.* 40, 1016–1030. doi: 10.3758/s13421-012-0204-6

Cohen, A. L., Sanborn, A. N., and Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychon. Bull. Rev.* 15, 692–712. doi: 10.3758/PBR.15.4.692

Dobbins, I. G., and Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: evidence for an item-based criterion placement heuristic. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 1186–1198. doi: 10.1037/0278-7393.31.6.1186

Dube, C., Rotello, C. M., and Heit, E. (2010). Assessing the belief bias effect with ROCs: it's a response bias effect. *Psychol. Rev.* 117, 831–863. doi: 10.1037/a0019634

Elqayam, S., and Evans, J. S. (2011). Subtracting 'ought' from 'is': descriptivism versus normativism in the study of the human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Heit, E., and Rotello, C. M. (2014). Traditional difference-score analyses of reasoning are flawed. *Cognition* 131, 75–91. doi: 10.1016/j.cognition.2013.12.003

Klauer, K. C., Musch, J., and Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychol. Rev.* 107, 852–884. doi: 10.1037/0033-295X.107.4.852

Morrell, H. E. R., Gaitan, S., and Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 1095–1110. doi: 10.1037/0278-7393.28.6.1095

Singmann, H., and Kellen, D. (2014). Concerns with the SDT approach to causal conditional reasoning: a comment on Trippas, Verde, Handley, Roser, McNair, and Evans (2014). *Front. Psychol.* 5:402. doi: 10.3389/fpsyg.2014.00402

Trippas, D., Handley, S. J., and Verde, M. F. (2013). The SDT model of belief bias: complexity, time, and cognitive ability mediate the effects of believability. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1393–1402. doi: 10.1037/a0032398

Trippas, D., Handley, S. J., and Verde, M. F. (2014b). Fluency and belief bias in deductive reasoning: new indices for old effects. *Front. Psychol.* 5:631. doi: 10.3389/fpsyg.2014.00631

Trippas, D., Verde, M. F., and Handley, S. J. (2014c). Using forced choice to test belief bias in syllogistic reasoning. *Cognition* 133, 586–600. doi: 10.1016/j.cognition.2014.08.009

Trippas, D., Verde, M. F., Handley, S. J., Roser, M. E., McNair, N. A., and Evans, J. S. (2014a). Modeling causal conditional reasoning data using SDT: caveats and new insights. *Front. Psychol.* 5:217. doi: 10.3389/fpsyg.2014.00217

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 08 October 2014; accepted: 05 February 2015; published online: 19 February 2015.

*Citation: Trippas D, Verde MF and Handley SJ (2015) Alleviating the concerns with the SDT approach to reasoning: reply to Singmann and Kellen (2014). Front. Psychol. **6**:184. doi: 10.3389/fpsyg.2015.00184*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 02 July 2014
doi: 10.3389/fpsyg.2014.00665

Heuristics and biases: interactions among numeracy, ability, and reflectiveness predict normative responding

Paul A. Klaczynski*

Decision Making and Development, School of Psychological Science, University of Northern Colorado, Greeley, CO, USA

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Adam Sanborn, University of Warwick, UK
Maggie E. Toplak, York University, Canada
Kinga Morsanyi, Queen's University Belfast, UK

***Correspondence:**
Paul A. Klaczynski, School of Psychological Science, College of Education and Behavioral Sciences, University of Northern Colorado, McKee Hall, Greeley, CO 80639, USA
e-mail: paul.klaczynski@unco.edu

In Stanovich's (2009a, 2011) dual-process theory, analytic processing occurs in the algorithmic and reflective minds. Thinking dispositions, indexes of reflective mind functioning, are believed to regulate operations at the algorithmic level, indexed by general cognitive ability. General limitations at the algorithmic level impose constraints on, and affect the adequacy of, specific strategies and abilities (e.g., numeracy). In a study of 216 undergraduates, the hypothesis that thinking dispositions and general ability moderate the relationship between numeracy (understanding of mathematical concepts and attention to numerical information) and normative responses on probabilistic heuristics and biases (HB) problems was tested. Although all three individual difference measures predicted normative responses, the numeracy-normative response association depended on thinking dispositions and general ability. Specifically, numeracy directly affected normative responding only at relatively high levels of thinking dispositions and general ability. At low levels of thinking dispositions, neither general ability nor numeric skills related to normative responses. Discussion focuses on the consistency of these findings with the hypothesis that the implementation of specific skills is constrained by limitations at both the reflective level and the algorithmic level, methodological limitations that prohibit definitive conclusions, and alternative explanations.

Keywords: normative, heuristics and biases, analytic processing, moderator effects, numeracy

INTRODUCTION

When the standards against which they are evaluated are traditional norms, performance on heuristics and biases (HB) tasks is often poor (Kahneman et al., 1982; Reyna and Brainerd, 1995; Stanovich, 1999). Underlying most views of the "normative/descriptive gap" (see Baron, 2008) is the assumption that rational thinking is "bounded" by information processing limitations (e.g., working memory, processing speed). In accord with this view, measured intelligence, generally assumed to index these processing limitations, relates positively to normative responses on several HB tasks. To the extent that measured intelligence accurately taps individual differences in cognitive capacity, these findings partially support the "bounded rationality" hypothesis. The general modesty of the correlations (*r*s range = 0.20–0.45; see Stanovich and West, 2008) implies, however, that considerable variance in responding cannot be easily attributed to computational limitations (see also Reyna, 2000).

Evidence that differences in general ability account for 20% (or less) of the variability in normative responses was at least partially responsible for research on the associations between responses and less "bounded" individual difference variables. Thus, in addition to research on specific intellectual competencies (e.g., inhibition; Markovits et al., 2009; De Neys, 2012; Markovits, 2013), the focus of numerous investigations has been the relationship between thinking dispositions (TD) and HB responses (e.g., Stanovich and West, 1998; see Stanovich, 2009b, 2012). Thinking dispositions—relatively malleable cognitive styles, beliefs, intellectual values, and motivations to manage cognitive resources (e.g., expending effort, guarding against impulsivity, valuing deliberate thinking, openness to using different strategies)—often account for variance in performance independently from general ability (Stanovich and West, 1998, 2000; Klaczynski and Lavallee, 2005; West et al., 2008; Toplak et al., 2011).

Research on TD and general ability (GA) has led to theoretical models that distinguish between two levels of analytic processing. The most common distinction in dual-process theories is between autonomous (or "Type I") processing and analytic (or "Type II") processing (e.g., Evans, 2009, 2011; Klaczynski, 2009; Barrouillet, 2011; Stanovich, 2011; Evans and Stanovich, 2013). Autonomous processing is triggered by task/situational factors, operates without conscious awareness and automatically activates situationally-relevant heuristics and other memories (e.g., procedural) that can serve as the basis for inferences and judgments. Analytic processing is conscious, deliberate, and cognitively demanding and is responsible for judging the adequacy of autonomously-produced representations and responses, determining whether to override autonomous processing, and engaging conscious reasoning and decision making abilities (see Stanovich, 1999, 2009a; Klaczynski, 2004; Evans, 2007). When predominant, analytic processing guides the selection and operation of the cognitive strategies and underlies complex reasoning and computations (Stanovich, 2011).

Stanovich (2004, 2009a; Stanovich and West, 2008; Stanovich et al., 2011) has proposed the analytic processes are best conceived as operating in two related "minds": The reflective mind and the algorithmic mind—hereafter referred to as the reflective and algorithmic levels. Reflective-level operations, generally indexed by measures of epistemic understanding and thinking dispositions, regulate or govern algorithmic-level activities and are therefore metacognitive in nature. The algorithmic level, most often indexed by measures of intelligence, comprises general cognitive competencies, information processing efficiency (e.g., working memory), reasoning abilities (inductive, deductive), and specific computational and logical rules, strategies, and abilities. This description suggests that the algorithmic level can be partitioned into (a) general abilities, resources, and limitations on processing efficiency and (b) specific abilities or "micro-strategies" (see Stanovich, 2009a, p. 71). General processing resources are superordinate to specific abilities in the sense that, in the absence of sufficient resources, even individuals who possess the abilities (e.g., numeracy, described subsequently) to solve particular problems will be incapable of fully utilizing those abilities and will therefore err in their attempts.

The conceptual relationships among the reflective level, general algorithmic-level resources, and specific algorithmic abilities can be summarized as follows. First, because the reflective level guides operations (e.g., specific strategy selection, computation monitoring, response evaluation) at the algorithmic level, it is superordinate to both general algorithmic resources and specific algorithmic skills. Second, despite being "subordinate," available algorithmic resources necessarily limit the efficiency of reflective-level functions. Third, the same algorithmic limitations impose constraints on the quality (e.g., complexity) and functionality of specific skills.

The present research was intended to provide a preliminary test of the model of analytic processing outlined above and examine the associations among thinking dispositions, general ability, and numeracy. Broadly defined, numeracy is set of specific algorithmic "micro-strategies" encompassing individuals' understanding of, and ability to assign meaning to, mathematical concepts (Nelson et al., 2008; Peters, 2012). Because numerous HB tasks require at least a minimal understanding of probabilities, numeracy is an algorithmic skill set with considerable promise for advancing our understanding of the processes underlying performance. Indeed, extant research indicates that numeracy is associated with general ability and explains variance on some HB tasks beyond that attributable to general ability and more specific aspects of algorithmic competence (e.g., inhibition; Peters et al., 2006; Nelson et al., 2008; Liberali et al., 2011; Toplak et al., 2011). Despite these findings, several hypotheses directly relevant to Stanovich's theory of analytic processing have not been examined.

Specifically, the view of Stanovich's theory espoused here, that reflective operations guide general algorithmic operations and that both reflective and algorithmic operations are important determinants of whether numeric skills are used to generate normative responses, implies specific conditions under which numeracy predicts normative responses on probabilistic tasks. Because reflective operations are critical to judging the adequacy of automatically-activated representations and responses, determining whether decoupling is necessary, understanding task requirements (e.g., whether problems require numeric computations), selecting specific algorithmic skills, monitoring computational operations, and evaluating outcomes, a first condition is adequate reflective-level functioning. A second necessary condition is the availability of sufficient general algorithmic capacity: Algorithmic resources (e.g., working memory) are required not only to perform reflective operations and sustain decoupled representations but also to effectively utilize numeric abilities and conduct computations (Stanovich and West, 2008). Thus, the effects of numeracy on responses should depend on (i.e., be moderated by) thinking dispositions and cognitive ability. This conjecture led to the hypotheses described below and depicted in **Figure 1**.

(1) *Inadequate reflective-level regulation*. Inadequacies at the reflective level should result in poor management of general algorithmic resources, little attention to representation quality or consideration of alternative representations, errors in specific ability selection, and little monitoring of algorithmic operations. Therefore, regardless of general ability, numeracy was not expected to relate to normative responding among participants with poorly developed thinking dispositions.

(2) *Inadequate general algorithmic resources*. Because algorithmic resources limit the efficiency of both reflective-level functions and numeric operations, participants low in general ability were expected to respond non-normatively—regardless of thinking dispositions and numeric ability.

(3a) *Low numeric ability*. Regardless of levels of reflective and algorithmic functioning, to perform well on probabilistic problems, individuals must have adequate numeric abilities. Those with poor numeric abilities were expected to respond non-normatively—regardless of reflective skills (TD) and algorithmic resources (GA).

(3b) *High numeric ability.* From the model described previously and the preceding hypotheses, it follows that, among participants high in numeric ability, only those who also have high levels of thinking dispositions *and* general intellectual ability would respond normatively[1].

The above predictions apply only to conflict problems—that is, problems wherein different responses are implied by task content (e.g., stereotype-activating information) and task structure (e.g., probability information). In contrast to *conflict* (i.e., CN) problems, on *no-conflict* (i.e., N-CN) problems, responses triggered automatically by task content are the same (i.e., normative) as responses based correct application of analytic abilities (De Neys,

[1]The theoretical speculations advanced here imply causal relationships among thinking dispositions, general ability, and numeracy. However, with the exception of the conflict/no-conflict problem comparisons, the research was correlational. In the Results, terms that connote causality are sometimes used because of space considerations and because such terms (e.g., "direct" effects) are used in discussions of moderation. Although the observed relationships may be causal, they should be interpreted cautiously and with consideration of alternative explanations (see Discussion).

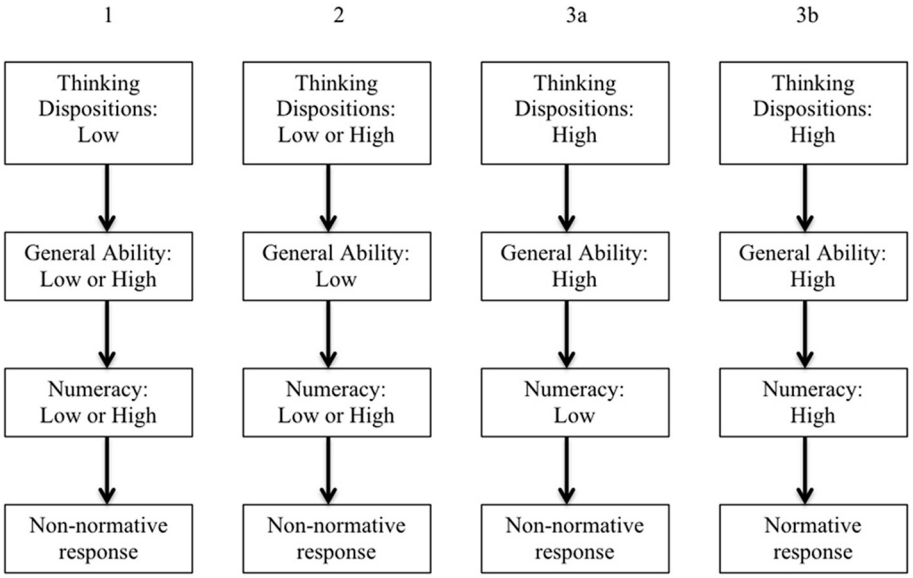


FIGURE 1 | Predicted relationships among thinking dispositions, general ability, numeracy, and normative responses on probabilistic HB tasks.

2012). Although responses on N-CN problems have been examined in some investigations (e.g., De Neys and Van Gelder, 2009; Thompson and Johnson, 2014; see also research on belief-biased reasoning; e.g., Evans et al., 1983), N-CN problems are often not examined in HB research. However, because normative responses should be considerably more frequent on N-CN problems than on CN problems and because N-CN responses should not be diagnostic of underlying processes, performance on N-CN problems should correlate with neither performance on CN problems nor the individual difference measures. Preliminary analyses were intended to explore these hypotheses for no-conflict problems (in a sense, the N-CN problems served as control problems; see De Neys, 2012).

METHODS

PARTICIPANTS

As part of a larger investigation, 219 undergraduates earned course credit for participating in single 60–80 min session (in groups of 4–8 students) during which they reported their verbal and quantitative SAT scores, completed measures of numeracy, general ability, and thinking dispositions, and responded to a battery of HB tasks.

MATERIALS

Thinking dispositions

The 52-item TD questionnaire, based on similar measures used by Stanovich and West (e.g., Stanovich and West, 1998, 2007) and Klaczynski (e.g., Klaczynski and Lavallee, 2005), contained five subscales (items were intermixed randomly). The 10-item *flexible thinking* scale measured willingness to take into account multiple perspectives and beliefs that complex decisions cannot be reduced to "either-or" choices (Macpherson and Stanovich, 2007). The 10-item *reflectiveness* vs. *intuition* scale assessed beliefs that logic and careful analysis leads to better decisions than reliance on intuitions (Epstein et al., 1995). The 12-item *need for cognition* scale measured valuation of intellectual challenges, complex thinking, and logical deliberation (see Cacioppo et al., 1996). The 14-item *impulsive decision making* scale tapped tendencies to make decisions "on the spur of the moment" (i.e., without considering consequences or alternatives) and believe that the best decisions are made quickly (see Patton et al., 1995). The 8-item *epistemic regulation* scale indexed understanding that belief conflicts can be resolved by considering the best available evidence (based on Kuhn, 2006 and Moshman, 2013). Participants responded to each item on a 6-point scale (1 = *strongly disagree*; 6 = *strongly agree*).

To reduce the number of analyses, a composite TD score was computed ($M = 161.68$, $SD = 13.85$). The composite was justified by the positive correlations among subscales (smallest $r = 0.25$) and the higher internal consistency ($\alpha = 0.78$) and stronger correlations with responses for the composite than for the subscales.

General ability

Both verbal ability and inductive reasoning ability were assessed. Verbal ability, best indexed by vocabulary, is among the foremost indicators of global and crystallized intelligence. Fluid intelligence, perhaps the best indicator of algorithmic-level functioning (Stanovich, 2009a,b), was indexed by scores on an inductive reasoning test.

Verbal ability. A 30-item vocabulary test ($M = 21.87$; $SD = 2.72$), based on the Shipley-2 vocabulary test (Shipley et al., 2010), was administered. Pilot testing indicated a correlation of 0.89 between the revised and the original tests. The Shipley-2 has excellent internal and test-retest reliability and relates moderately/strongly to academic achievement, general intelligence, and other indexes of crystallized intelligence (Prokosch et al., 2005;

Kaya et al., 2012). On each item, a target word (e.g., jocose) was followed by four options (e.g., humorous, paltry, fervid, plain). Correct responses required selecting the word with same meaning as the target. Three minutes were given to complete as many items as possible.

Inductive ability. A 20-item inductive reasoning test ($M = 10.75$; $SD = 1.72$) was administered. Items were selected after removing the easiest and most difficult items from the PMA Letter Sets test (Thurstone, 1962). In pilot testing, the original PMA and the reduced version were correlated highly ($r = 0.84$). Scores on the original test and shortened versions of the test correlate well with general intelligence and other indexes of fluid intelligence (Hertzog and Bleckley, 2001; Colom et al., 2007). From five sets of four letters (e.g., ACDE, MOPQ, FGIJ, DFGH, TVWX), participants indicated the set that did not belong with the other sets (e.g., FGIJ) and completed as many items as they could in 12 min.

A composite ability score was analyzed for several reasons. First, inductive and verbal scores correlated moderately ($r = 0.47$; Kaya et al., 2012, reported a similar correlation). Second, scores on the two measures related similarly to normative responses. Third, the combined ability score correlated better (see **Table 2**) with normative responses than inductive ability (*r*s ranged from 0.21 to 0.28) or verbal ability (*r*s ranged from 0.22 to 0.27).

Numeracy

Participants completed a 20-item objective numeracy test ($\alpha = 0.82$; $M = 11.39$, $SD = 3.53$). Objective numeracy tests (in contrast to subjective tests) contain items that measure basic probability skills, such as those involved in converting ratios to percentages (and vice versa) and analyzing fractions (e.g., 2/20 vs. 3/40) to determine relative probabilities. The numeracy test (available from the author) was similar to the tests used by Peters et al. (2006), Nelson et al. (2008), and Liberali et al. (2011) and an included items from (or adapted from) Lipkus et al. (2001), Garfield (2003), Irwin and Irwin (2005), and Klaczynski and Amsel (2014).

Each item included a problem that required understanding a probabilistic concept and selecting, from 3–5 response options, the correct solution (e.g., from a list of 20 names, the chances a randomly selected name would begin an "A"; the probability that a randomly selected person would be a doctor who also enjoys hiking in a group of 100 people with three doctors and eight people who enjoy hiking). The predictive value and validity of the test were established in two developmental studies of responses on HB problems similar to those described subsequently. In both studies, numeracy increased with age and accounted for more variance in normative responding than age or ability. Using a similar measure, Klaczynski and Amsel (2014) found that numeracy predicted differences on probabilistic reasoning tasks better than age or nationality (Chinese or American).

Heuristics and biases tasks

Given the definition of numeracy given previously, numeracy should be a better predictor of normative responses on *probabilistic* HB problems than of normative responses on non-probabilistic problems. The battery, presented in one of four randomly determined orders and mixed with problems from a larger study (order was not related to responses on any HB task or to any of the individual difference measures), included eight base rate neglect (BR), eight law of large numbers (LLN), eight ratio bias (RB), and eight covariation judgment (COV) problems. For each of task (i.e., BR, LLN, RB, COV), there were four conflict (CN) problems and four no-conflict (N-CN) problems. On both the conflict and no-conflict versions of each task, normative scores could range from 0 to 4; mean proportions of normative responses are presented in the Results to increase the ease of comparing the findings with other research. Examples of conflict and no-conflict versions of each task are presented in the Supplementary material[2].

Base rate neglect problems. Each problem intended to elicit base rate neglect contained two types of information: (1) Base rate data indicating the number of people in each of two groups and (2) descriptions of individual "targets" that were consistent with stereotypes associated with one group (e.g., knitting, gardening). On CN problems, target descriptions "pulled" for responses based on group stereotypes and the base rates (e.g., 125 17-year-olds and 25 50-year-olds) pulled for the normative response that targets were not likely to be members of the stereotyped groups. The stereotypes thus cued responses that conflicted with normative responses. The target descriptions in the N-CN problems were identical to those in the CN problems; however, on the N-CN problems the base rates (e.g., 25 17-year-olds and 125 50-year-olds) indicated that targets were likely in the stereotyped group. Normative responses were thus cued by both the stereotypes and the base rates (see also De Neys and Glumicic, 2008).

On each problem, participants judged target group membership on 4-point scales (e.g., 1 = *Very likely to be 17 years old*; 2 = *Somewhat likely 17 years old*; 3 = *Somewhat likely to be 50 years old;* 4 = *Very likely to be 50 years old*; reversed for half the problems). Consistent with previous studies (e.g., Toplak et al., 2014), responses on the CN problems were considered normative (scored "1") when participants rated that targets as unlikely or very unlikely to be in the stereotyped group and responses on the N-CN problems were scored normative when participants rated targets as likely or very likely to be in the stereotyped group.

Law of large numbers. Adapted from Fong et al. (1986), Stanovich and West (1998), and Klaczynski (2001), these problems involved making decisions after reviewing arguments founded on large evidential samples and arguments based on small samples of personal and relatively vivid evidence. On CN problems, large sample arguments supported one decision and small sample arguments supported a different decision. On the N-CN problems, the large sample and small sample arguments supported the same decision. On four problems (two CN, two N-CN), the large sample arguments were presented before the

[2]In a larger investigation, numeracy was only related weakly to responses on non-probabilistic problems. Despite GA and TD correlations to responses similar to those reported here, the TD × GA × Numeracy interaction was not significant; instead, the TD × GA interaction was a significant predictor of responses.

small sample arguments. On the other four problems (two CN, two N-CN), the small sample arguments were presented first.

Participants indicated the decision they judged best on 4-point scales (1 = "Decision 'A' is a much better decision"; 2 = "Decision 'A' is a better decision"; 3 = "Decision 'B' is a better decision"; 4 = "Decision B is a much better decision," where "Decision B" indicated preference for the large sample argument). For half the problems, the rating scale was reversed and later recoded; consequently, on both the CN and N-CN problems, ratings of 3 and 4 reflected greater reliance on the large sample arguments. Following Stanovich and West (1998), Klaczynski (2001), and Toplak et al. (2007), ratings ≥3 were considered normative and assigned scores of 1.

Ratio bias. On the RB problems (Denes-Raj and Epstein, 1994), participants judged whether targets (e.g., winning lottery tickets) were more likely if a person selected from a relatively large numerator/large denominator sample (e.g., nine winning tickets in 100 total tickets) or a relatively small numerator/small denominator sample (e.g., one winning ticket in 10 total tickets). The RB effect occurs when individuals believe that targets are more likely from relatively large samples than from relatively small samples. Reyna and Brainerd (2008) distinguished between heuristic RB problems (i.e., identical probabilities in the two samples) and non-optimal RB problems (i.e., probabilities favor the smaller sample). Although the RB effect has been reported on both heuristic and non-optimal problems, non-optimal problems were used in the present research because the normative response (e.g., on CN problems, targets were more likely from the smaller sample) was more similar to normative responses on the other tasks than was the normative response on heuristic problems (i.e., neither sample is more likely to yield a target).

On each problem, the absolute number of targets (i.e., numerators) and the total (i.e., targets plus non-targets; denominators) was higher in the large sample than in the small sample. On CN problems, target probability was higher in the smaller sample. By contrast, on N-CN problems, the absolute numbers of targets and the probabilities of targets were higher in the larger samples: Similar to the N-CN contingency detection problems (described next), normative selections could be based on calculating and comparing ratios or simply comparing numerators. On two CN and two N-CN problems, the small sample response was presented before by the large sample response; on the other CN and N-CN problems, the larger sample option was presented before the small sample option. A third option (that target probability was the same in the two samples) was always presented last. Participants judged which, if either, sample was more likely to yield a target (e.g., winning lottery ticket). Judgments were normative (scored "1") when the small sample was selected on the CN problems and the large sample was selected on the N-CN problems.

Covariation judgment problems. Based on Wasserman et al. (1990) and modeled after the problems in Stanovich and West (1998), Klaczynski (2001), and De Neys and Van Gelder (2009), each problem described a hypothetical investigation of a potentially causal relationship between two variables. Descriptions were accompanied by 2 × 2 contingency tables summarizing the results (i.e., numbers of cases) in each of the four cells: (putative) cause-present/effect-present, cause-absent/effect-present, cause-absent/effect-present, and cause-absent/effect-absent (labeled the A–D cells; Wasserman et al., 1990). Relationship strength can be determined by computing phi (φ) or comparing conditional probabilities [A/(A + B) − C/(C + D)], although less precise ratio comparisons yield relationships in the same direction as φ. When Cell A is clearly larger and more salient than Cell B (and Cell C), adults often adopt the simple strategy of comparing numbers of cases in Cell A with the numbers of Cell B (or Cell C; see Alloy and Tabachnik, 1984; Maldonado et al., 2006). As discussed by fuzzy-trace theorists, this numerosity bias is similar to that found on RB problems (see Reyna and Brainerd, 2008).

On the CN problems, the absolute numbers in Cell A (e.g., 35) were greater than the numbers in cell B (e.g., 26) and cell C (e.g., 27), but the φ coefficients were negative (in this example, Cell D was 11). Thus, judgments based on comparing Cell A with Cell B or Cell C conflicted with judgments based on computing φ or comparing ratios. On the N-CN problems, the absolute numbers in Cell A (e.g., 37) were also greater than the numbers in Cells B (e.g., 15) and C (e.g., 23), but the φ coefficients were positive (e.g., 18 in Cell B). Thus, normative solutions could be based on computing conditional probabilities, comparing ratios, or simply comparing Cell A with Cell B or Cell C.

Participants judged relationship strength on 5-point scales (1 = *strong negative relationship*; 5 = *strong positive relationship*; reversed for two CN and two NC problems). After recoding problems with reversed rating scales, responses were judged normative (scored "1") when participants indicated that the correlations were negative (i.e., ratings <3) on the CN problems and positive on the N-CN problems (i.e., ratings >3).

PROCEDURE

The ability measures, because they were timed, were always administered before the other measures. For about half of the participants, the HB battery was presented next, followed by the thinking dispositions questionnaire and the numeracy measure. For the remaining participants, presentation order was the thinking dispositions questionnaire, numeracy test, and HB battery. Order was not significantly related to either normative responses or individual difference variables (largest $r = 0.11$).

RESULTS

CONFLICT AND NO-CONFLICT PROBLEMS

To examine whether normative responses were more frequent on N-CN problems than on CN problems, a multivariate analysis of variance, with normative scores on the four tasks as dependent variables and problem type (CN or N-CN) as a within-subjects variable, was conducted. The anticipated main effect of problem type was significant, $F_{(1,\ 215)} = 1617.26$, $p < 0.001$, $\eta_p^2 = 0.88$: On each task, normative responses were more frequent on N-CN problems than on CN problems, smallest $F_{(1,\ 215)} = 295.17$, $p < 0.001$, $\eta_p^2 = 0.60$. Mean proportions of normative responses on the conflict and no-conflict problems are presented in **Table 1**.

CORRELATIONS BETWEEN NORMATIVE RESPONSES AND PREDICTORS

The next analyses were intended to determine whether no-conflict scores on the four tasks were related to each other, conflict scores, and the individual difference measures (i.e., TD, GA, and numeracy). With the exception of negative correlations between scores on the NC-LLN and CN-COV problems and between scores on the NC-RB and NC-COV problems, no correlations between no-conflict scores on the different tasks or between responses on no-conflict and conflict problems were significant (see **Table 2**). Similarly, no correlations between the individual difference variables and no-conflict scores were significant (largest $r = 0.11$). Next, the correlations among responses on the conflict versions of the tasks and the correlations among the hypothesized predictors were examined. As expected, and consistent with prior research (Stanovich and West, 1998; Klaczynski, 2001; Chiesi et al., 2011), responses on the conflict versions of each task correlated positively (see **Table 2**). The predictors were also significantly related (TD-ability = 0.19, $p < 0.01$; TD-numeracy = 0.22, $p < 0.01$; ability-numeracy = 0.31, $p < 0.001$). SAT scores also related to TD, ability, and numeracy ($rs = 0.27$, 0.22, 0.25, respectively, all $ps < 0.01$). However, when they were significant, the relationships between SAT scores and normative responses were weak relative to the correlations between normative responses and the other predictors (see **Table 3**).

More central to the goals of this investigation were the correlations between conflict responses and the hypothesized predictors. Note that, although the relationships between normative responses and *interactions* between predictors (e.g., Numeracy × Ability) are not typically examined in HB research (see, however, Stanovich and West, 2008; Chiesi et al., 2011; Handley et al., 2011), the study's hypotheses required analyses of these relationships. That is, positive correlations between the TD × Ability × Numeracy interaction and normative responses would be consistent—and thus provide initial support for—the speculation that effects associated with numeracy are constrained by ability and TD.

The correlations of the individual predictors and the predictor interaction terms (computed by standardizing and then multiplying TD, ability, and numeracy scores) to responses on each task and a composite score (normative responses on each task summed and divided by four) are presented in **Table 3**. TD, ability, and numeracy correlated positively with individual task scores and composite scores, supporting the hypothesis that each variable would predict responses. Of the two-way interactions, the Ability × Numeracy interaction correlated positively with the individual task scores and composite scores. More important, however, were the significant correlations of the TD × Ability × Numeracy interaction to individual task scores and composite scores. As noted above, these particular correlations are consistent with the speculation that the "effects" of numeracy on responses were at least partially constrained by ability and TD. Although promising, the findings from this analysis represent only a first step toward testing the hypothesis. An important second step entailed determining whether the three-way interaction explained variance in normative responses beyond that associated with the individual predictors and the two-way predictor interactions.

PREDICTING NORMATIVE RESPONSES

In and of themselves, the correlational findings do not indicate whether TD constrained the numeracy-response relationships or, alternatively, whether TD constrained the ability-response relationships. To reduce the number of additional analyses, subsequent analyses focused on composite scores. This focus is justified by the significant relationships among individual task scores, a principal components factor analysis that yielded a single score with an eigenvalue > 1 (54.18% of the variance among scores; smallest loading = 0.69), and the finding that results for the individual tasks closely paralleled the results from the analyses of the composite[3].

[3]In subsequent analyses of composite scores, similar results obtained when factor scores were analyzed.

Table 1 | Mean proportions (and *SD*s) of normative responses on the conflict and no-conflict problems.

Task	Conflict	No conflict
Base rate	0.49 (0.24)	0.87 (0.21)
Law of large numbers	0.41 (0.23)	0.92 (0.16)
Ratio bias	0.36 (0.21)	0.90 (0.19)
Covariation	0.34 (0.19)	0.88 (0.19)

Table 2 | Correlations between responses on the conflict and no-conflict problems.

	2	3	4	5	6	7	8
1. CN: BR	0.34[c]	0.34[c]	0.26[c]	−0.02	−0.12	−0.06	0.09
2. CN: LLN		0.36[c]	0.31[c]	0.09	0.11	0.02	−0.03
3. CN: RB			0.40[c]	−0.01	−0.09	−0.01	−0.02
4. CN: COV				0.05	−0.14[a]	0.09	−0.11
5. N-CN: BR					0.03	−0.09	−0.18[b]
6. N-CN: LLN						−0.06	0.02
7. C-CN: RB							0.07
8. N-CN: COV							

[a]$p < 0.05$; [b]$p < 0.01$; [c]$p < 0.001$.

Table 3 | Correlations between predictors and responses on the conflict problems.

	BR	LLN	RB	COV	Comp.
SAT	0.10	0.11	0.15[a]	0.15[a]	0.18[b]
TD	0.27[c]	0.25[c]	0.28[c]	0.32[c]	0.36[c]
Ability	0.28[c]	0.28[c]	0.31[c]	0.30[c]	0.38[c]
Numeracy	0.30[c]	0.25[c]	0.28[b]	0.31[c]	0.39[c]
TD × Ability	0.05	0.09	0.06	0.02	0.08
TD × Numeracy	0.04	0.14[a]	0.12	0.06	0.10
Ability × Numeracy	0.17[b]	0.16[a]	0.19[b]	0.21[b]	0.25[c]
TD × Ability × Numeracy	0.28[c]	0.26[c]	0.26[c]	0.32[c]	0.36[c]

[a]$p < 0.05$; [b]$p < 0.01$; [c]$p < 0.001$.

To determine (a) which predictors accounted for unique variance in normative responses and (b) whether the predictor interaction terms accounted for variance in composite scores beyond the variance associated with the individual predictors, a hierarchical multiple regression analysis was conducted on composite scores. SAT-Math scores were entered at the first step and TD, GA, and numeracy were entered at the second step. To determine whether they accounted for additional variance, the two-way interaction terms were entered at the third step and the three-way interaction term was entered at the final step. Significant contributions of the TD × Numeracy and GA × Numeracy interactions would suggest that numeracy moderated the relationships of TD and GA to normative responses and a significant contribution of the TD × GA × Numeracy interaction would imply that the numeracy-normative response relationship depended on both TD and GA[4].

Results from the final step, and incremental variance explained by the predictors at each step, are presented in **Table 3**. In total, the predictors and interaction terms accounted for 35.9% of the variance in composite scores. TD, ability, and numeracy were significant independent predictors, as were the GA × Numeracy and the TD × GA × Numeracy interactions. The significant predictive value of these interactions implies that the effects of ability, numeracy, and TD were less straightforward than implied by the significant beta values of the individual predictors. The three-way interaction, which contributed an additional 2.1% of variance beyond that explained by the other predictors, is particularly important because it implies that the numeracy-normative response relationship depended on GA and TD. Unfortunately, the regression results provide little information regarding the specific nature of the interactive relationships and thus do not fully address the investigation's central hypothesis. Although consistent with the Hypotheses (3a) and (3b), the significant predictive value of the three-way interaction does not indicate that the numeracy-normative association differed for low and high TD participants whose general abilities were low or high and therefore is insufficient evidence for conclusions regarding the constraining effects of TD and GA on the numeracy-normative response association. Consequently, an alternative approach was needed to determine whether the relationship between numeracy and normative responses depended on whether thinking dispositions and general ability were high or low.

ABILITY AND THINKING DISPOSITIONS AS MODERATORS OF THE NUMERACY-RESPONSE ASSOCIATION

The hypothesis that the numeracy-response relationship would be significant only if TD and general ability were relatively high is a moderation hypothesis. To test the speculation that numeracy differences depended on both ability and thinking dispositions, Hayes' (2012; for related discussions, see Shrout and Bolger, 2002; Preacher et al., 2007; Hayes, 2013) SPSS macro and, specifically, "process model 3" was used to conduct a "moderated moderation" analysis. In brief, the process macro uses ordinary least squares regression to estimate the coefficients for each predictor and their interactions. Process model 3 is useful in determining the significance of the interactions between and among an independent variable and two moderators. Results indicated whether effects related to numeracy depended on GA and TD and whether the numeracy-composite relationship was significant only when GA and TD were relatively high. As suggested by the foregoing regression analyses, support for the hypothesis was contingent on the significance of the three-way interaction (i.e., Numeracy × GA × TD)[5].

By default, Hayes' (2012) macro constructs three levels (subsequently referred to as "low," "moderate," and "high"; levels are centered around the means; i.e., the mean and ± 1 SD from the mean) for the IV and each moderator. If the three-way interaction is significant, these levels are used to examine the significance of the interaction between numeracy and ability at each level of the moderator (TD). At least in a general sense, the analysis parallels a 3 (numeracy) × 3 (ability) × 3 (TD) analysis of variance. However, unlike analysis of variance approaches, but consistent with current approaches to moderation and mediation (Preacher et al., 2007), bootstrapping procedures are used to obtain 95% confidence intervals. Confidence intervals provided a basis for estimating whether, at each ability level within each TD level, numeracy was significantly related to composite scores. Effects were considered significant when confidence intervals did not contain zero (Hayes, 2012). In the results presented below, LLCI and ULCI refer to lower level and upper level confidence interval, respectively.

To test the hypothesis that numeracy would "directly" affect responses only when TD and GA were high, numeracy was entered as the "independent" variable, TD was entered as a one moderator, and ability was entered a second moderator. SAT-MATH scores and a composite N-CN score were entered as covariates. As in the regression analysis, the covariates, numeracy, GA, TD, and their interactions accounted for 36% of the variance in composite scores, $F_{(9,\ 206)} = 12.20$, $p < 0.0001$. TD ($\beta = 0.0023$, $t = 2.98$, $p = 0.0032$, LLCI/ULCI = 0.0008/0.0038), ability ($\beta = 0.0092$, $t = 3.47$, $p = 0.0006$; LLCI/ULCI = 0.0042/0.0152), and numeracy ($\beta = 0.0092$, $t = 2.81$, $p = 0.0054$; LLCI/ULCI = 0.0027/0.0156) were significant predictors (neither covariate was a significant

[4]No interaction that included total SAT scores or SAT-MATH scores related to, or predicted, composite scores. However, because they related to composite scores, SAT scores were included in subsequent analyses as covariates.

[5]Similar results obtained from a 2 (TD group) × 2 (ability group) × 2 (numeracy group) ANOVA on composite scores. In the low TD group, no effects related to numeracy were significant ($ps > 0.20$). In the high TD group, the Ability × Numeracy interaction was significant, $F_{(1,\ 103)} = 16.71$, $p < 0.001$, $\eta_p^2 = 0.07$. When TD was high, but ability was low, scores did not differ in the by numeracy group, $F < 1$. However, when TD and ability were high, the high numeracy group performed better than the low numeracy group, $F_{(1,\ 61)} = 35.83$, $p < 0.001$, $\eta_p^2 = 0.37$. Similar results obtained when Bayes factor—indicating the likelihood that the high and low numeracy groups differed—was computed at each TD and GA level (with r was set at 0.50; see http://pcl.missouri.edu/bayesfactor). Comparisons between the high and low numeracy groups for (a) low TD/low GA participants, (b) low TD/high GA participants, (c) high TD/low GA participants, and (d) high TD/high GA participant yielded Bayes factors of 1.113, 0.417, 1.923, and 7.283, respectively (the final factor is considered moderate/strong).

predictor, $ts < 1$). The Ability × Numeracy interaction ($\beta = 0.0021$, $t = 2.29$, $p = 0.0178$; LLCI/ULCI = 0.0004/0.0038) and the TD × Ability × Numeracy ($\beta = 0.0001$, $t = 2.15$, $p = 0.0322$; ULCI/LLCI = 0.0000/0.0002) interactions were significant. The three-way interaction indicated that the effects related to numeracy differed by levels of thinking dispositions and ability[6].

The results presented in **Table 4** show the effects of GA and numeracy at each TD level. As expected, the Numeracy × Ability interaction was not significant at the lowest TD level. Indeed, when TD low, the numeracy-response association was not significant at any ability level. By contrast, at moderate and high levels of TD, the Numeracy × Ability interaction was significant. The additional results shown in the table revealed that, when TD was moderate or high, numeracy directly affected normative responses only if GA was also moderate or high. These findings, depicted in **Figure 2**, support the general hypothesis that TD and GA constrained the effects of numeracy on responding to probabilistic HB tasks.

[6]Hayes (2012) refers to Process 3 as a "moderated moderation" analysis, intended to determine whether the effects of an independent variable interact with the effects of two other variables (moderators). Although the decision to enter TD and GA as moderators and numeracy as the IV was theoretical, the analysis is nonetheless analogous to an analysis of variance (see Footnote 5) with three levels for each "IV." As such, the three-way interaction was significant regardless of which variables were entered as moderators and which was entered as the IV. For instance, with numeracy was the IV, GA as the first moderator, and TD as the second moderator, the variance explained was identical. The primary difference is that, instead of presenting the GA × Numeracy interaction (and simple effects of numeracy) within each TD level, this alternative analysis indicated whether the TD × Numeracy interaction was significant at each GA level and, within GA levels, the direct affects of numeracy when TD was low, moderate, and high. However, in contrast to the findings presented here, the TD × Numeracy interaction was significant only when GA was high. Otherwise, the results of the follow-up analyses were analogous to those in **Table 5**: Numeracy directly affected responses when GA was moderate and high and when TD was also moderate and high.

Table 4 | Hierarchical multiple regression analysis on composite scores (β and *t*-values from final step).

Predictors	$R\Delta^2$	$F\Delta$	*B*	β	*t*
SAT	0.03	6.14[a]	0.00	−0.02	<1
TD, ability, numeracy	0.27	27.72[c]			
TD			0.00	0.20	3.12[b]
GA			0.01	0.24	3.66[c]
Numeracy			0.01	0.20	3.19[b]
Two-way interactions	0.04	3.66[a]			
TD × Ability			0.00	0.01	<1
TD × Numeracy			0.01	0.05	<1
Ability × Numeracy			0.03	0.18	2.90[b]
Numeracy × Ability × TD	0.02	7.11[c]	0.02	0.17	2.68[b]

[a]$p < 0.05$; [b]$p < 0.01$; [c]$p = 0.001$.

DISCUSSION

This study showed that normative responses on no-conflict problems are typically related to neither responses on conflict problems nor thinking dispositions, general ability, or numeracy. By contrast, normative responses on conflict problems related positively to all three individual difference variables. After accounting for variance attributable to thinking dispositions, general ability, and numeracy entered separately, the Thinking Disposition × General Ability × Numeracy interaction accounted for additional variance in normative responses on the conflict problems.

Perhaps the most important contribution of the present research are the findings bearing on hypotheses based on Stanovich's (2009a, 2011) theory of analytic processing. As anticipated by Hypotheses (1) and (2), when TD was low—regardless of whether general ability was low, moderate, or high—and when GA was low—regardless of whether thinking dispositions were low, moderate, or high —numeracy was unrelated to normative responses. Although based on correlational data, these preliminary findings are consistent with the proposed relationship between the reflective and algorithmic levels. Deficiencies at the reflective level appear to limit the efficacy of algorithmic functions. Thus, even the most intellectually able (regardless of numeric ability) solved few probabilistic HB problems correctly when their epistemic beliefs and thinking dispositions were poorly calibrated. Conversely, algorithmic limitations appear to

Table 5 | Moderated mediation results: effects of numeracy on normative responding by TD level and ability level (within TD levels).

Numeracy × ability	Predicting composite normative responses			
	Estimate	*t*	LLCI	ULCI
Low TD	0.0003	<1	−0.00021	0.0028
Ability				
Low	0.0058	>1	−0.10057	0.0174
Moderate	0.0071	1.57	−0.5771a	0.0161
High	0.0085	1.18	−0.1885a	0.0226
Moderate TD	0.0021	2.39[a]	0.0004	0.0038
Ability				
Low	0.0012	<1	−0.0088	0.0113
Moderate	0.0092	2.81[a]	0.0027	0.0156
High	0.0171	4.09[b]	0.0089	0.0254
High TD	0.0038	3.35[b]	0.0016	0.0061
Ability				
Low	−0.0034	<1	−0.0180	0.0113
Moderate	0.0112	2.55[a]	0.0025	0.0199
High	0.0258	5.55[b]	0.0166	0.0350

Note. Numeracy × Ability = Numeracy × Ability interaction at each TD level. Within TD levels, significance of numeracy at low, moderate, high ability levels. Ability and TD levels are derived from means and ± one SD (TD ± 13.85; Ability ± 3.83) from the respective means. LLCI and ULCI = 95% bias corrected lower lever confidence interval and upper level confidence interval, respectively (5000 bootstrap samples). [a]$p < 0.05$; [b]$p < 0.001$.

constrain the efficacy of reflective functions: Participants at the highest level of reflective functioning (regardless of numeric ability) performed little better than those at the lowest TD level when they lacked the cognitive resources to conduct reflective operations (e.g., selecting appropriate micro-strategies or mindware, evaluating task representations) and perform correct computations.

Among the most novel contributions of this research, however, were those pertaining to Hypothesis (3b). Consistent with expectations, when TD was moderate-high *and* ability was moderate-high, numeracy associated positively with normative responding. The effects of numeracy were thus moderated by both thinking dispositions and ability. These findings support the position that relatively high levels of reflective and general algorithmic functioning are both necessary for numeracy to influence responding, at least on probabilistic tasks. As indicated in **Figure 2**, when they lacked *either* the requisite thinking dispositions *or* general intellectual competencies, highly numeric individuals were no more likely than less numeric individuals to respond normatively.

To a greater extent than prior research, these findings support the perspective previously outlined on Stanovich's (2009a, 2011) theory of analytic processing. First, the findings were not limited to a single task but extended across four probabilistic reasoning tasks. Second, few investigations have entailed examinations of the interactive effects of thinking dispositions, general ability, and specific abilities (micro-strategies or mindware) on reasoning. Third, the moderated moderation analytic approach afforded a more precise exploration of the hypothesized relationships than other approaches (e.g., ANOVAs based on median split-created groups). Finally, the results not only implicated numeracy as an important contributor to probabilistic reasoning but also provided theoretically-consistent evidence relevant to the conditions under which numeracy predicts normative responding: When TD *and* GA are both fairly high (note that the precise meaning of "moderate" and "high" TD and GA is relative to the population studied and depends on the measures used to assess these constructs).

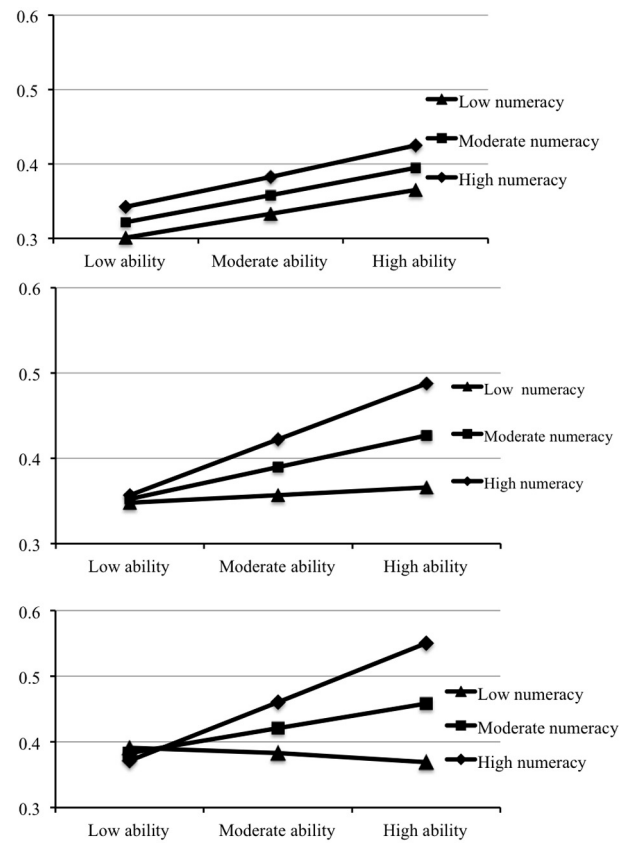


FIGURE 2 | Moderated moderation results: effects of numeracy levels and ability level on normative responding at low (upper graph), moderate (middle graph), and high (bottom graph) TD levels.

The research presented here was concerned with processes that ensue *after* conflict detection and *after* decisions to attempt overriding autonomously-triggered responses with responses based on analytic processing. In the dual-process theory advocated by Evans and Stanovich (e.g., Evans, 2007, 2008, 2012; Stanovich, 2009a,b, 2012; Reyna and Brainerd, 2011; Evans and Stanovich, 2013), rapid processing of problem content activates potential responses. These autonomous responses are not necessarily inadequate or non-normative (Handley et al., 2011; Thompson and Johnson, 2014); instead, they are accompanied by varying "feelings of rightness" (Thompson, 2009; Thompson et al., 2013). Notably, the findings of Handley, Thompson, and colleagues, indicating that normative responses are sometimes automatically activated, provide additional weight to cautionary notes to guard against assuming that analytic processing necessarily underlies normative responses (e.g., Klaczynski, 2001; Reyna et al., 2003; Elqayam and Evans, 2011; Evans, 2011; Reyna and Brainerd, 2011; Stanovich et al., 2011). In the present work, normative responses may sometimes have been activated automatically, a possibility that might partially explain why thinking dispositions, general ability, and numeracy accounted for only 36% of the response variance. As implied below, measures of "feeling of rightness" and inhibition would likely have explained additional variance.

The stronger the "feelings of rightness" elicited by automatic responses, the lower the probability that reasoners will attempt to replace these responses with consciously deliberated answers (Thompson and Morsanyi, 2012; Thompson et al., 2013). The model tested here is therefore likely more relevant to autonomous responses associated with weak "rightness feelings" (or sensing "something fishy" about intuitive responses; De Neys, 2012, p. 31). At a minimum level, the decision to judge the sufficiency of the intuitive responses that trigger weak "feelings of rightness" is a metacognitive, reflective process. However, to further engage analytic processes and fully evaluate automatic responses, both reflective operations and algorithmic resources are required (the latter to compare intuitive responses against responses based on careful deliberation and to internalized standards; see also Moshman, 1998). If an automatically-activated response is deemed inadequate (e.g., inaccurate and/or insufficiently precise), reflective abilities again come into play to assess task requirements, select the appropriate algorithmic skills, and judge the outcomes of implementing those skills. Algorithmic resources are, of course, not only necessary to carry out these

procedures and implement specific reasoning, decision making, and computational skills, but also to suppress initial responses and inhibit interference from potentially misleading beliefs activated by task content (e.g., stereotypes) or by the intuitive responses themselves.

To summarize, metacognitive operations at the reflective level determine whether override should be attempted (Klaczynski, 2004; Thompson, 2009; Evans, 2010). Following this decision, the generation of decoupled representations depends on reflective functioning (e.g., recognition of task requirements/structure) which, in turn, is dependent on general algorithmic resources and specific experiences and skills. After such representations are generated, the appropriate mindware (e.g., numeracy)—if available—must be selected (Stanovich, 2012). Even if available, correct strategy/skill selection does not guarantee that implementation will be effective. Inability to sustain generated representations and inhibit autonomous responses (effortful processes requiring both algorithmic resources and reflective dispositions; see Stanovich and West, 2008) can lead to interference from non-essential task contents and implementation errors (see also the discussion of "levels of rationality" in Reyna et al., 2003). Clearly, as anticipated by the arguments and supported by the evidence proffered by Reyna et al. (2003) and others (e.g., Evans, 2011; Stanovich et al., 2011; Klaczynski, 2013), attempts to override responses based on autonomous processing are neither invariably successful nor invariably lead to normative responses.

By themselves, neither algorithmic capabilities (including specific mindware) nor competence at the reflective level sufficed to produce normative responses. In Stanovich's theory, the reflective-algorithmic relationship is reciprocal because reflective operations are necessarily constrained by available resources (Stanovich and West, 2008). Thus, even those at the highest general ability and numeracy levels typically gave non-normative responses when their reflective dispositions and skills were poor (see also Overton, 1990; Amsel et al., 2008; Chiesi et al., 2011; Ricco and Overton, 2011; Morsanyi and Handley, 2013). Several reflective-level difficulties, such as failures to accurately assess task requirements, attend to numerical information in accurate representations, select appropriate computational skills, monitor numeric functions and outputs, or equate *subjectively-adequate* responses with normative responses, could have led to non-normative responses. Conversely, even participants at the highest levels of reflectivity and numeracy typically gave non-normative responses if their general ability scores were low. Lacking the requisite resources to implement and monitor their numeric skills while maintaining decoupled representations (see Stanovich et al., 2011, 2012; Stanovich, 2012), these individuals performed no better than those at low levels of reflective functioning and numeracy.

The findings support the theory of analytic processing proposed by Stanovich (2009a, 2011) and implicate numeracy as a specific algorithmic skill likely to further our understanding of the processes underlying performance on HB tasks. Research on the role of instructions in reasoning is also consistent, and can be interpreted from the perspective of, Stanovich's theory. Evidence from several reports indicates that reliance on heuristics decreases and normative responses increase when participants are instructed to think logically (e.g., Denes-Raj and Epstein, 1994). Recent findings (e.g., Macpherson and Stanovich, 2007; Evans et al., 2010; Handley et al., 2011; Morsanyi and Handley, 2012; Morsanyi et al., 2012) have further demonstrated that such instructions improve responding primarily among high ability participants and that, in the absence of such instructions, general ability is unrelated to responding on some tasks. If conceived as externally-imposed surrogates for well-calibrated thinking dispositions—or as cues to engage in reflective-level operations—logic instructions should only benefit those with sufficient algorithmic capacity to not only keep the instructions in mind but also construct accurate representations and conduct the relevant computations. Just as it constrains reflective-level functioning, general ability limits the efficacy of logic instructions.

Despite evidence consistent with the view that a function of the reflective level is to select, guide, and monitor algorithmic operations and that algorithmic limitations constrain not only these reflective operations but also the implementation of specific abilities, there are reasons to guard against interpreting the current findings as definitive support for this theoretic position. Specifically, the correlational nature of the study prohibits the conclusions that thinking dispositions *constrained* the functioning of general ability and that limitations in general ability *constrained* numeric operations (see Footnotes 1 and 6). For instance, the hypothesized relationship between thinking dispositions and general ability is reciprocal; however, it was not possible to examine directly bidirectional (or unidirectional) causal relationships in the present work. Even if the causal relationships operate as hypothesized on probabilistic tasks, the model does not explain findings that, on some HB tasks, (a) thinking dispositions sometimes predict performance but general ability does not, (b) general ability sometimes predicts performance but thinking dispositions do not, and (c) neither thinking dispositions nor general ability relate positively to performance (e.g., Klaczynski, 2000; Stanovich and West, 2008; Thompson and Johnson, 2014). These mixed and sometimes null findings may, to some extent, be attributable to the fact that measures of general ability and thinking dispositions are imperfect indexes of algorithmic and reflective functioning. Replications of, for instance, research on myside biases that utilizes more specific (and/or more extensive) measures of algorithmic (e.g., inhibition) and reflective (e.g., metacognitive monitoring) processes would likely contribute valuable insights toward explaining these findings.

Another issue is that the individual differences measures accounted for only 36% of response variance. One reason for this, alluded to earlier, is that normative responses are sometimes activated automatically. In such instances, complete engagement of analytic resources is not always necessary (reasoners may even forgo checks of response override when automatic normative responses are accompanied by strong feelings of rightness). An expansion of this account may also help explain the unexplained variance: When initial responses prompt attempts to override and to construct decoupled representations, it is conceivable that the process of assessing task requirements automatically activates normative responses. That is, the effort that goes into override and/or decoupling may be sufficient to trigger normative responses. In such cases, algorithmic resources would be taxed

little (see Thompson and Johnson, 2014) and reflective operations would be relatively limited (e.g., monitoring computation quality would neither be necessary nor possible). This account, however, awaits empirical testing.

Nonetheless, at least on probabilistic reasoning tasks, the combination of well-calibrated beliefs and intellectual dispositions with moderate-high cognitive ability may well lead to normative responses *if* specific micro-strategies or mindware (e.g., numeracy) are available. The findings thus lend additional substance to recent discussions of dual-process theories, support the distinction between the reflective and algorithmic levels of analytic processing, and contribute new data to the growing literature on numeracy. Even so, additional research examining the interactions among thinking dispositions, general ability, and specific abilities is clearly needed. In conducting these investigations, theory-driven moderation (and mediation) analyses will likely yield results more informative than those based on less precise analyses (e.g., ANOVA). When coupled with findings from experimental research, our understanding of the processes that underlie judgments, reasoning, and decisions will likely improve considerably. Arguments over whether responses judged normative should be considered prescriptive can be better addressed empirically. As an example, if general abilities are subordinated to thinking dispositions/epistemic regulation and the latter can be acquired through formal and informal tuition—and if some specific algorithmic abilities are educable—then the possibility the reducing the gap between traditional norms ("what ought") and actual behavior ("what is") remains open (for discussion and alternative perspectives, see Elqayam and Evans, 2011).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00665/abstract

REFERENCES

Alloy, L. B., and Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychol. Rev.* 91, 112–149. doi: 10.1037/0033-295X.91.1.112

Amsel, E., Klaczynski, P. A., Johnston, A., Bench, S., Close, J., Sadler, E., et al. (2008). A dual-process account of the development of scientific reasoning: the nature and development of metacognitive intercession skills. *Cogn. Dev.* 23, 452–471. doi: 10.1016/j.cogdev.2008.09.002

Baron, J. (2008). *Thinking and Deciding, 4th Edn.* New York, NY: Cambridge University Press.

Barrouillet, P. (2011). Dual-process theories and cognitive development: advances and challenges. *Dev. Rev.* 31, 79–85. doi: 10.1016/j.dr.2011.07.002

Cacioppo, J. T., Petty, R. E., Feinstein, J., and Jarvis, W. (1996). Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychol. Bull.* 119, 197–253. doi: 10.1037/0033-2909.119.2.197

Chiesi, F., Primi, C., and Morsanyi, K. (2011). Developmental changes in probabilistic reasoning: the role of cognitive capacity, instructions, thinking styles, and relevant knowledge. *Think. Reason.* 17, 315–350. doi: 10.1080/13546783.2011.598401

Colom, R., Escorial, S., Shih, P. C., and Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Pers. Individ. Dif.* 42, 1503–1514. doi: 10.1016/j.paid.2006.10.023

Denes-Raj, V., and Epstein, E. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *J. Pers. Soc. Psychol.* 66, 819–829. doi: 10.1037/0022-3514.66.5.819

De Neys, W. (2012). Bias and conflict: a case for logical intuitions. *Perspect. Psychol. Sci.* 7, 28–38. doi: 10.1177/1745691611429354

De Neys, W., and Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition* 106, 1248–1299. doi: 10.1016/j.cognition.2007.06.002

De Neys, W., and Van Gelder, E. (2009). Logic and belief across the lifespan: the rise and fall of belief inhibition during syllogistic reasoning. *Dev. Sci.* 12, 123–130. doi: 10.1111/j.1467-7687.2008.00746.x

Elqayam, S., and Evans, J. St., B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–290. doi: 10.1017/S0140525X1100001X

Epstein, S., Pacini, R., Denes-Raj, V., and Heier, H. (1995). *Individual Differences in Rational and Analytical Information Processing.* Amherst, MA: University of Massachusetts.

Evans, J. St. B. T. (2007). On the resolution of conflict in dual-process theories of reasoning. *Think. Reason.* 13, 321–329. doi: 10.1080/13546780601008825

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annu. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Evans, J. St. B. T. (2009). "How many dual-process theories do we need: One, two or many?" in *In Two Minds: Dual Processes and Beyond,* eds J. St. B. T. Evans and K. Frankish (Oxford: Oxford University Press), 31–54.

Evans, J. St. B. T. (2010). *Thinking Twice: Two Minds in One Brain*. Oxford: Oxford University Press.

Evans, J. St. B. T. (2011). Dual process theories of reasoning: contemporary issues and developmental applications. *Dev. Rev.* 31, 86–102. doi: 10.1016/j.dr.2011.07.007

Evans, J. St. B. T. (2012). Spot the difference: distinguishing between two kinds of processing. *Mind Soc.* 11, 121–131. doi: 10.1007/s11299-012-0104-2

Evans, J. St. B. T., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Mem. Cogn.* 11, 295–306. doi: 10.3758/BF03196976

Evans, J. St. B. T., Handley S. J., Neilens, H., and Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *Q. J. Exp. Psychol.* 63, 892–909. doi: 10.1080/17470210903111821

Evans, J. St. B., T., and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Fong, G. T., Krantz, D. H., and Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cogn. Psychol.* 18, 253–292. doi: 10.1016/0010-0285(86)90001-0

Garfield, J. B. (2003). Assessing statistical reasoning. *Stat. Educ. Res. J.* 2, 22–38. Available online at: http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(1).pdf

Handley, S. J., Newstead, S. E., and Trippas, D. (2011). Logic, beliefs, and instruction: a test of the default interventionist account of belief bias. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 28–43. doi: 10.1037/a0021098

Hayes, A. F. (2012). *PROCESS: a Versatile Computational Tool for Observed Variable Mediation, Moderation, and Conditional Process Modeling. White Paper.* The Ohio State University. Available online at: http://www.afhayes.com/public/process2012.pdf (Accessed February, 2012).

Hayes, A. F. (2013). *SPSS, SAS, and Mplus Macros and Code.* Available online at: http://afhayes.com/introduction-to-mediation-moderation-and-conditional-process-analysis.html (Accessed: September 24, 2013).

Hertzog, C., and Bleckley, M. K. (2001). Age differences in the structure of intelligence: Influences of information processing speed. *Intelligence* 29, 191–217. doi: 10.1016/S0160-2896(00)00050-7

Irwin, K. C., and Irwin, R. J. (2005). Assessing development in numeracy of students from different socio-economic areas: a Rasch analysis of three fundamental tasks. *Educ. Stud. Math.* 58, 283–229 doi: 10.1007/s10649-005-6425-x

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511809477

Kaya, F., Delen, E., and Bulut, O. (2012). Test review: Shipley-2 manual. *J. Psychoeduc. Assess.* 30, 593–597. doi: 10.1177/0734282912440852

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: a two-process approach to adolescent cognition. *Child Dev.* 71, 1347–1366. doi: 10.1111/1467-8624.00232

Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision making. *Child Dev.* 72, 844–861. doi: 10.1111/1467-8624.00319

Klaczynski, P. A. (2004). "A dual-process model of adolescent development: implications for decision making, reasoning, and identity," in *Advances in Child Development and Behavior,* Vol. 31, ed R. V. Kail (San Diego, CA: Academic Press), 73–123.

Klaczynski, P. A. (2009). "Cognitive and social cognitive development: dual-process research and theory," in *Two Minds: Psychological and Philosophical Theories of Dual Processing,* eds J. B. St. T. Evans and K. Frankish (Oxford: Oxford University Press), 265–292.

Klaczynski, P. A. (2013). "Culture and the development of heuristics and biases: implications for developmental dual-process theories," in *The Development in Thinking and Reasoning,* eds P. Barrouillet and C. Gauffroy (London: Psychology Press), 150–192.

Klaczynski, P. A., and Amsel, E. A. (2014). *Numeracy and Age Predict Chinese and American Children's Heuristics and Biases*. Greeley, CO: University of Northern Colorado.

Klaczynski, P. A., and Lavallee, K. L. (2005). Domain-specific identity, epistemic regulation, and intellectual ability as predictors of belief-based reasoning: a dual-process perspective. *J. Exp. Child Psychol.* 92, 1-24. doi: 10.1016/j.jecp.2005.05.001

Kuhn, D. (2006). Do cognitive changes accompany developments in the adolescent brain? *Perspect. Psychol. Sci.* 1, 59–67. doi: 10.1111/j.1745-6924.2006.t01-2-.x

Liberali, J. M., Reyna, V. E., Furlan, S., Stein, L. M., and Pardo, S. T. (2011). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *J. Behav. Decis. Mak.* 25, 361–381. doi: 10.1002/bdm.752

Lipkus, I. M., Samsa, G., and Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Med. Decis. Making* 21, 37–44. doi: 10.1177/0272989X0102100105

Macpherson, R., and Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learn. Individ. Differ.* 17, 115–127. doi: 10.1016/j.lindif.2007.05.003

Maldonado, A., Jimenez, G., Herrera, A., Perales, J. C., and Catena, A. (2006). Inattentional blindness for negative relationships in human causal learning. *Q. J. Exp. Psychol.* 59, 457–470. doi: 10.1080/02724980443000854

Markovits, H. (2013). "The development of abstract conditional reasoning," in *The Development of Thinking and Reasoning,* eds P. Barrouillet and C. Gauffroy (New York, NY: Psychology Press), 71–91.

Markovits, H., Saelen, C., and Forgues, H. L. (2009). An inverse belief–bias effect: More evidence for the role of inhibitory processes in logical reasoning. *Exp. Psychol.* 56, 112–120. doi: 10.1027/1618-3169.56.2.112

Morsanyi, K., and Handley, S. (2013). "Heuristics and biases: insights from developmental studies," in *The Development of Thinking and Reasoning,* eds P. Barrouillet and C. Gauffroy (New York, NY: Psychology Press), 122–149.

Morsanyi, K., and Handley, S. J. (2012). Logic feels so good—I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 596–616. doi: 10.1037/a0026099

Morsanyi, K., Primi, C., Chiesi, F., and Handley, S. J. (2012). The effects and side-effects of statistics education: psychology students' (mis-)conceptions of probability. *Contemp. Educ. Psychol.* 34, 210–220. doi: 10.1016/j.cedpsych.2009.05.001

Moshman, D. (1998). "Cognitive development beyond childhood," in *Handbook of child Psychology:* Vol. 2. *Cognition, Perception, and Languagem, 5th Edn.,* eds D. Kuhn and R. Siegler, Series ed W. Damon (New York, NY: Wiley), 947–978.

Moshman, D. (2013). "Epistemic cognition and development," in *The Development of Thinking and Reasoning,* eds P. Barrouillet and C. Gauffroy (New York, NY: Psychology Press), 13–149.

Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I., and Peters, E. (2008). Clinical implications of numeracy: theory and practice. *Ann. Behav. Med.* 35, 261–274. doi: 10.1007/s12160-008-9037-8

Overton, W. F. (1990). "Competence and procedures: constraints on the development of logical reasoning," in *Reasoning, Necessity, and Logic: Developmental Perspectives,* ed W. F. Overton (Hillsdale, NJ: Erlbaum), 1–32.

Patton, J., Stanford, M., and Barratt, E. (1995). Factor structure of the Barratt Impulsiveness Scale. *J. Clin. Psychol.* 51, 768–774.

Peters, E. (2012). Beyond comprehension: the role of numeracy in judgments and decisions. *Curr. Dir. Psychol. Sci.* 21, 31–35. doi: 10.1177/0963721411429960

Peters, E., Västfjäll, D., Slovic, P., Mertz, C., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. *Psychol. Sci.* 17, 407–413. doi: 10.1111/j.1467-9280.2006.01720.x

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing moderated mediation hypotheses: theory, methods, and prescriptions. *Multivariate Behav. Res.* 42, 185–227. doi: 10.1080/00273170701341316

Prokosch, M. D., Yeo, R. A., and Miller, G. F. (2005). Intelligence tests with higher g-loadings show higher correlations with body symmetry: evidence for a general fitness factor mediated by developmental stability. *Intelligence* 33, 203–213. doi: 10.1016/j.intell.2004.07.007

Reyna, V. F. (2000). Data, development, and dual processes in rationality. *Behav. Brain Sci.* 23, 694–695. doi: 10.1017/S0140525X0054343X

Reyna, V. F., and Brainerd, C. J. (1995). Fuzzy-trace theory: an interim synthesis. *Learn. Individ. Differ.* 7, 1–75. doi: 10.1016/1041-6080(95)90031-4

Reyna, V. F., and Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learn. Individ. Differ.* 18, 89–107. doi: 10.1016/j.lindif.2007.03.011

Reyna, V. F., and Brainerd, C. J. (2011). Dual processes in decision making and developmental neuroscience: a fuzzy-trace model. *Dev. Rev.* 31, 180–206. doi: 10.1016/j.dr.2011.07.004

Reyna, V. F., Lloyd, F. J., and Brainerd, C. J. (2003). "Memory, development, and rationality: an integrative theory of judgment and decision making," in *Emerging Perspectives on Judgment and Decision Research,* eds S. L. Schneider and J. Shanteau (New York, NY: Cambridge University Press), 201–245. doi: 10.1017/CBO9780511609978.009

Ricco, R. B., and Overton, W. F. (2011). Dual systems competence ?-? procedural processing: a relational developmental systems approach to reasoning. *Dev. Rev.* 31, 119–150. doi: 10.1016/j.dr.2011.07.005

Shipley, W. C., Gruber, C., Martin, T., and Klein, A. M. (2010). *Shipley Institute of Living Scale, 2nd Edn.* Los Angeles, CA: Western Psychological Services.

Shrout, P. E., and Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol. Methods* 7, 422–445. doi: 10.1037/1082-989X.7.4.422

Stanovich, K., West, R. F., and Toplak, M. E. (2011). The complexity of developmental predictions from dual process models. *Dev. Rev.* 31, 103–118. doi: 10.1016/j.dr.2011.07.003

Stanovich, K. E. (1999). *Who is Rational? Studies of Individual Differences in Reasoning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning the Age of Darwin*. Chicago, IL: Chicago University Press. doi: 10.7208/chicago/9780226771199.001.0001

Stanovich, K. E. (2009a). "Is it time for a tri-process theory. Distinguishing the reflective and the algorithmic mind," in *In Two Minds: Dual Processes and Beyond,* eds J. St. B. T. Evans and K. Frankish (Oxford: Oxford University Press), 55–88.

Stanovich, K. E. (2009b). *What Intelligence Tests Miss. The Psychology of Rational Thought.* New Haven, CT: Yale University Press.

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. New York, NY: Oxford University Press.

Stanovich, K. E. (2012). "On the distinction between rationality and intelligence: implications for understanding individual differences in reasoning," in *The Oxford Handbook of Thinking and Reasoning,* eds K. J. Holyoak and R. G. Morrison (Oxford: Oxford University Press), 433–455.

Stanovich, K. E., and West, R. F. (1998). Individual differences in rational thought. *J. Exp. Psychol. Gen.* 127, 161–188. doi: 10.1037/0096-3445.127.2.161

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate. *Behav. Brain Sci.* 23, 645–726. doi: 10.1017/S0140525X00003435

Stanovich, K. E., and West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Think. Reason.* 13, 225–247. doi: 10.1080/13546780600780796

Stanovich, K. E., and West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *J. Pers. Soc. Psychol.* 94, 672–695. doi: 10.1037/0022-3514.94.4.672

Stanovich, K. E., West, R. F., and Toplak, M. E. (2012). "Intelligence and rationality," in *Cambridge Handbook of Intelligence, 3rd Edn.,* eds R. Sternberg and S. B. Kaufman (Cambridge: Cambridge University Press), 784–826.

Thompson, V. A. (2009). "Dual process theories: a metacognitive perspective," in *In Two Minds: Dual Processes and Beyond,* eds J. Evans and K. Frankish (Oxford: Oxford University Press). 171–196.

Thompson, V. A., and Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking, *Think. Reason.* 20, 215–244. doi: 10.1080/13546783.2013.869763

Thompson, V. A., and Morsanyi, K. (2012). Analytic thinking: do you feel it? *Mind Soc* 11, 93–105. doi: 10.1007/s11299-012-0100-6

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., et al. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition* 128, 237–251. doi: 10.1016/j.cognition.2012.09.012

Thurstone, T. G. (1962). *Primary Mental Abilities For Grades 9–12*. Chicago, IL: Science Research Associates.

Toplak, M. E., Liu, E., Macpherson, R., Toneatto, T., and Stanovich, K. E. (2007). The reasoning skills and thinking dispositions of problem gamblers: a dual-process taxonomy. *J. Behav. Decis. Mak.* 20, 103–124. doi: 10.1002/bdm.544

Toplak, M. E., Stanovich, K. E., and West, R. F. (2014). Rational thinking and cognitive sophistication: development, cognitive abilities, and thinking dispositions. *Dev. Psychol.* 50, 1037–1048. doi: 10.1037/a0034910

Toplak, M. E., West, R. F., and Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Mem. Cognit.* 39,1275–1289. doi: 10.3758/s13421-011-0104-1

Wasserman, E. A., Dorner, W. W., and Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *J. Exp. Psychol. Learn. Mem. Cogn.* 16, 509–521. doi: 10.1037/0278-7393.16.3.509

West, R. F., Toplak, M. E., and Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions. *J. Educ. Psychol.* 100, 930–941. doi: 10.1037/a0012842

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 09 June 2014; published online: 02 July 2014.
*Citation: Klaczynski PA (2014) Heuristics and biases: interactions among numeracy, ability, and reflectiveness predict normative responding. Front. Psychol. **5**:665. doi: 10.3389/fpsyg.2014.00665*
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 18 June 2014
doi: 10.3389/fpsyg.2014.00581

Rationality: a social-epistemology perspective

Sylvia Wenmackers[1†], Danny E. P. Vanpoucke[2] and Igor Douven[1]*

[1] *Faculty of Philosophy, University of Groningen, Groningen, Netherlands*
[2] *Center for Molecular Modeling, Ghent University, Ghent, Belgium*

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
David E. Over, Durham University, UK
Rainer Hegselmann, Bayreuther Forschungszentrums Modellierung und Simulation sozioökonomischer Phänomene, Germany

****Correspondence:***
Igor Douven, Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, Netherlands
e-mail: i.e.j.douven@rug.nl

[†] *Address from October 2014: Institute of Philosophy, University of Leuven, Leuven, Belgium*

Both in philosophy and in psychology, human rationality has traditionally been studied from an "individualistic" perspective. Recently, social epistemologists have drawn attention to the fact that epistemic interactions among agents also give rise to important questions concerning rationality. In previous work, we have used a formal model to assess the risk that a particular type of social-epistemic interactions lead agents with initially consistent belief states into inconsistent belief states. Here, we continue this work by investigating the dynamics to which these interactions may give rise in the population as a whole.

Keywords: social epistemology, rationality, computer simulations, opinion dynamics, beliefs, theory, inconsistency, probability

1. INTRODUCTION

This paper aims to show the importance of a social perspective in the study of human rationality. While, as will be seen, the work we present relies on computer simulations, we believe it may inspire further empirical research by social scientists. Computer simulations, such as those to be presented, form a bridge between normative models and descriptive results. The simulations depend on a theoretical model with various parameters. Some combinations of the parameters may be optimal for the attainment of one or more norms, whereas other combinations of parameters may give a good approximation to an epistemic group of real people. If the model parameters can be linked to variables in the real world, this may enable us to give practical advice for increasing rationality in social settings.

In previous work, we studied a formal model of a type of epistemic interactions in which agents whose belief states are in some sense close together compromise by settling on a kind of "averaging" belief state. We showed that compromising in this way carries the risk of leading agents with initially consistent belief states to become inconsistent. Although it was shown in the same paper how this risk could be minimized, it might nonetheless be considered as a reason for banishing the designated kind of interactions. Here, we continue the previous work by investigating the dynamics of a population as a whole to which epistemic compromising may give rise. We pay special attention to the conditions under which such compromising may lead to a consensus among the members of a population. This is intended to shed new light on the question of whether it is at all rational to interact epistemically in the said kind of way.

2. THEORETICAL BACKGROUND

In their study of human rationality, philosophers as well as psychologists of reasoning have tended to focus on individual thinkers in isolation from their social environment. Which beliefs an individual ought to hold and how an individual ought to change his beliefs have traditionally been regarded as questions that are independent of which beliefs other individuals hold or how other individuals change their beliefs. This is at least somewhat surprising, given that we are so obviously members of a community of individuals who pursue by and large the same epistemic goals, who frequently engage in common activities to gather new evidence, who constantly exchange information, who often (have to) rely on the words of others, who regularly seek each other's advice in epistemic matters, and who sometimes put great effort into trying to influence one another's opinions. In fact, we see these kinds of behavior not just in everyday life, but also, and even especially, in the practice of science, which many regard as producing the—in some sense—best and most valuable knowledge. Doubtlessly, there are more and less rational ways of engaging in these various activities, and it would seem part of the business of philosophy, as well as of that of psychology, to sort out which are which.

At least in philosophy, there is a growing awareness that the general neglect of the group level in studying human rationality has created a serious gap in our understanding indeed, and philosophers have begun to correct this lacuna[1]. Their efforts

[1] Psychologists have recently begun to explore connections between rationality and argumentation, which can also be regarded as an appreciation of the fact that the social bears on questions of rationality. For a particularly noteworthy

have given rise to a field now commonly known as "social epistemology" (Goldman, 1999). Questions addressed so far by social epistemologists concern the possibility of testimonial justification (in particular, the question of whether we are justified in holding a belief on the basis of another person's testimony; see Douven and Cuypers, 2009, Fricker, 1987, and Lackey, 1999), the rationality (or otherwise) of aligning our opinions on a given matter with those of experts on the matter (Gaifman, 1986; van Fraassen, 1989, Ch. 8; Goldman, 2001), and the effect on our beliefs that the discovery of peer disagreement should have (that is, the question of whether we can rationally stick with our belief after the discovery that someone who we regard as a peer holds a contrary belief; see, for instance, Douven, 2009, 2010, and Elga, 2007).

The popularity of social epistemology being on the rise, it is easily missed that we still do not know whether, on balance, the possible benefits of social-epistemic interactions outweigh their possible costs. Indeed, various philosophers and also sociologists have enthusiastically reported about the "wisdom of the crowds," in the context of which it has been asserted that the aggregated opinions of a group of laypeople is often closer to the truth than the opinions of individual experts (Surowiecki, 2004). Such assertions might make one forget that crowds can be wildly erratic and irrational, too. We know how the crowd responded when, in his *Sportpalast* speech in February 1943, the Nazi minister of propaganda Joseph Goebbels asked whether it wanted total war. There was little wisdom in that response[2].

In trying to give cost–benefit analyses of diverse types of social epistemic interactions, and also for related purposes, a number of social epistemologists have recently started using computer simulations for studying communities of epistemically interacting artificial agents, where the agents typically adapt their beliefs (fully or partially) on the basis of information about the beliefs of other agents in the community. It has been argued that, insofar as these methods capture central aspects of the epistemic interactions between *real* agents, they give important information about the conduciveness of these interactions to the achievement of our epistemic goals as well as about the costs that may come with the interactions.

By far the most research on rationality is concerned either with developing an experimentally informed descriptive model of actual human thinking or with developing a theoretically-oriented normative model of idealized human thinking. We present a study that nominally falls in the first category, in that we study opinion dynamics with the help of computer experiments concerning epistemically interacting agents. But it would be more accurate to say that our study falls somewhere on the continuum between descriptive and normative work. The agents that we model are inspired by particular aspects of human thinking (such as the observation that humans have opinions on multiple topics, some of which are logically independent, and some of which are logically connected) and human epistemic interaction (most notably, that in practice we allow others' beliefs to influence our own as well as try to make our beliefs influence those of others), but—as will emerge—they clearly lack other characteristics of human thinking. So, it is not a purely descriptive study. The agents in the simulations do follow a prescribed way of revising their opinions and never fail to adhere to it, but they are still non-ideal thinkers; for example, they may come to believe an inconsistency without realizing this. Hence, it is not a purely normative model either.

As mentioned in the introduction, we here continue work that we have started elsewhere (Douven, 2010; Douven and Riegler, 2010, and especially Wenmackers et al., 2012). Specifically, we present a formal model for studying a community of agents that update their belief states by "averaging" (in a certain well-defined sense) over the belief states of agents that are close enough to their own belief state (where "close enough" will also receive a precise definition). In Wenmackers et al. (2012), we studied the question of how probable it is that averaging (in the designated way, yet to be specified) over others' belief states leads one into a state of inconsistency. Here, we investigate the opinion dynamics in a more global way: we consider the entire epistemic space (not just those situations in which some or all of the agents arrive at an inconsistency) and we do not restrict the dynamics to a single step (although it turns out that for the examples we consider, all of the dynamics plays out in just two steps). This approach gives us new insights into the previously obtained results and it allows us to visualize the process that the community as a whole undergoes as a result of the updates by its members. As stated in the introduction, we will be especially interested in the conditions under which the social-updating process leads to a consensus among the members of the community (including consensus on the inconsistent theory)[3]. We will give a brief summary at the end of each section. For a quick overview of the article, the reader may skip forward to these paragraphs of key points.

3. THE MODEL

The most widely known formal model for studying the effects of epistemic interactions on the belief states of individual agents

contribution in this vein, see Mercier and Sperber (2011), which argues that reasoning evolved primarily for argumentative purposes. See in this connection various of the contributions to the 2012 *Thinking & Reasoning* special issue on argumentation.

[2]See Andler (2012) for a critical discussion of the wisdom of the crowds idea.

[3]In our previous work, we concentrated on the possibility of ending up at the inconsistent theory, because believing a contradiction is generally considered as irrational. As a referee remarked, if the entire epistemic community ends up at the tautological theory, which represents a complete lack of knowledge about the world, this may also be considered as a vicious result—although not irrational *per se*. Since the current study does not focus on a particular result, but represents the dynamics in general, our results concerning the probabilities of ending up at the inconsistent theory are equally informative about the probabilities of ending up at the tautological theory. Specifically, for reasons of symmetry, the probabilities of arriving at a consensus on the tautological theory are identical to those for arriving at the inconsistent theory. In our previous work (Wenmackers et al., 2012), we only calculated the probability for one agent or the entire community to update to the inconsistent theory starting from a population in which no agent held this opinion, but we would have obtained the same probability values for one agent or all agents to update to the tautological theory starting from a population in which no agent held this opinion. Still, this inversion is less well motivated: starting out with the tautological theory is not necessarily a bad thing; some agents in the community may initially lack any evidence about the world and try to arrive at a more informative theory through social updating.

is probably the model developed in Hegselmann and Krause (2002, 2005, 2006), which now generally goes by the name of "Hegselmann–Krause model" ("HK model," for short). This model has received attention from researchers from various quarters, including philosophers, mathematicians, social scientists, and physicists (see, e.g., Deffuant et al., 2000; Dittmer, 2001; Weisbuch et al., 2002). It has also been used mainly to investigate descriptive questions, such as the question under which conditions the opinions of interacting agents are likely to polarize and under which conditions these opinions are likely to converge, but it has been used to investigate some normative issues as well (see Riegler and Douven, 2009; Douven, 2010). We consider a variant of the HK model. First, we present our general framework. In the course of this section, we present two examples. (Some readers may find it beneficial to consult the examples 3.1 and 3.2 prior to reading the more abstract setup.)

The basic version of the HK model assumes communities of agents that are trying to determine the value τ of some unspecified parameter by repeatedly and simultaneously averaging over the opinions of those agents that are within their so-called Bounded Confidence Interval (BCI)[4]. One agent is in a second agent's BCI—or, as we shall sometimes say, following Douven (2010), is a second agent's (epistemic) peer—precisely if the absolute difference between their opinions about the value of τ does not exceed some given threshold value ϵ. Hegselmann and Krause also study a model in which the agents take into account evidence about τ that they receive "directly from the world." More exactly, in this model the opinion of agent x_i after the $(u+1)$-th update is given by

$$x_i(u+1) = \alpha \frac{1}{|X_i(u)|} \sum_{j \in X_i(u)} x_j(u) + (1-\alpha)\tau, \qquad \text{(HK)}$$

where $x_i(u)$ is the opinion of agent x_i after the u-th update, whose peers (agents within the BCI after the u-th update) form the set $X_i(u) := \{j\colon |x_i(u) - x_j(u)| \leqslant \epsilon\}$, and $\alpha \in [0, 1]$ is the relative importance of the social-updating process as compared to evidence-gathering. In the basic version of the HK model, without evidence-gathering, $\alpha = 1$.

It is a limitation of the HK model that it considers only agents whose belief states consist, at any given point in time, of just one value. In Riegler and Douven (2009), an extension of the HK model was proposed that allows agents to have richer belief states in that they have beliefs on different aspects of the world. In other words, each agent holds a theory about the world, where a theory consists of a set of propositions expressible in the agent's language. A theory may be consistent or inconsistent: if no world can satisfy all the agent's beliefs—for instance, as when an agent believes that snow is white and also believes that snow is not white—then the agent holds an inconsistent theory about the world; otherwise the theory is consistent. Note that consistency does not guarantee truth: it may happen that some world or worlds satisfy all the agent's belief, but that the actual world does not. However, inconsistency does guarantee falsity: if a theory is true of no world—no world satisfies all of the agent's beliefs—then a fortiori it is not true of the actual world. Agents' belief states are supposed to be closed under (classical) logical derivability, meaning that any proposition expressible in the agent's language that follows logically from the agent's theory ipso facto belongs to that theory. As a result, the theory an agent holds can be represented by the strongest proposition it implies.

Given M atomic propositions, there are $w_M = 2^M$ possible worlds that we can distinguish between. In turn, this means that there are $t_M = 2^{w_M}$ theories about the world, exactly one of which represents the inconsistent theory, in which all the possible worlds have been ruled out by the agent. There is also exactly one tautology, the theory in which all possible worlds are left as epistemic possibilities for the agent. Note also that, by assuming some ordering of the possible worlds, the belief state of each agent can be represented by a bit string, where a 1-bit (0-bit) at the n-th location indicates that world number n (in the given ordering of worlds) is deemed possible (impossible) by the agent[5].

In this model, agents revise their theory of the world by taking into account the theories held by certain other agents in the community, comparable to how the agents in the HK model update. However, now the BCI is defined in a slightly more complicated way. To quantify the distance between two theories, the so-called Hamming distance δ between the corresponding bit strings is used: this distance is given by the number of locations in which these strings differ. The BCI is then defined by placing a threshold value D for δ, meaning that in updating the agents take into account the belief state of another agent if, and only if, the Hamming distance between (the bit string representing) the agent's own theory and (the bit string representing) the other agent's theory is smaller than or equal to D. An example may help to make this less abstract.

Example 3.1. Consider an interpreted propositional language $\mathcal{L}$ with just two atomic propositions, p, expressing that snow is white, and q, expressing that grass is green. Then there are 2^2 possible worlds: the world in which p and q both hold, the world in which p holds but q does not, the world in which q holds but p does not, and the world in which neither p nor q holds. Let these worlds be ordered in this way, so that the world in which both p and q hold is world number 1, and so on. Then the 16 theories that can be formulated in $\mathcal{L}$ can be coded as 4-digit strings. For example, the string 1111 codes the tautology: the actual world corresponds to one of the four possible worlds; the string 0000 codes the inconsistent theory: the actual world corresponds to none of the four possible worlds; and 1100 codes

[4]To forestall misunderstanding, it is worth mentioning that the word "Bounded" in "Bounded Confidence Interval" refers to the fact that the confidence intervals in the HK model have a lower and upper bound; in particular, the word is not meant to suggest any connection with Simon's notion of bounded rationality (see, e.g., Simon, 1955). The closest connection in the psychological literature is with the notion of confirmation bias, inasmuch as the BCI encompasses those agents whose opinions could be said to confirm to some extent one's own opinion.

[5]To avoid later disappointment, we note already at this juncture that, while we are introducing a general framework for representing theories, our own later investigations of this framework will focus on the $M = 1$ case.

the theory according to which snow is white and grass may or may not be green. Finally, if one agent holds the theory 1100 and another agent holds the theory 1001 (the theory according to which the world is such that *either* snow is white and grass is green *or* snow is not white and grass is not green), then the Hamming distance δ between their theories (that is, between the bit strings representing these theories) equals 2, given that they differ in the second and fourth bit and coincide otherwise. $\Diamond$

The update rule for theories in this model—so, basically the analog of (HK)—is a bitwise operation in two steps: (1) averaging and (2) rounding. In step (1), for each bit of the theory, a straight average is taken of the corresponding bit of those agents that are within the agent's BCI (note that this includes the agent himself). In general, the result is a value in the interval [0, 1] rather than just a 0 or 1. Hence the need for step (2): in case the average is greater than $1/2$, the corresponding bit is updated to 1; in case the average is less than $1/2$, the corresponding bit is updated to 0; and in case the average is exactly equal to $1/2$, the corresponding bit keeps its initial value.

More formally, the n-th bit of the bit string representation of agent x_i's belief state after the $(u+1)$-th update as determined by the extended HK update rule is

$$x_i(u+1)[n] = \begin{cases} 1 & \text{if } \frac{1}{|X_i(u)|} \sum_{j \in X_i(u)} x_j(u)[n] > \frac{1}{2}, \\ 0 & \text{if } \frac{1}{|X_i(u)|} \sum_{j \in X_i(u)} x_j(u)[n] < \frac{1}{2}, \\ x_i(u)[n] & \text{otherwise,} \end{cases} \tag{EHK}$$

with the set of peers of agent i after the u-th update now $X_i(u) := \{j\colon \delta(x_i(u), x_j(u)) \leqslant D\}$. Actually, in Riegler and Douven (2009) the agents also obtained evidence from the world, more or less as in one of the versions of the HK model. However, in our Wenmackers et al. (2012) we considered only the more basic (EHK), as we will do here. In Wenmackers et al. (2012) and also in the present paper, it is assumed that the agents update their beliefs simultaneously and repeatedly, at discrete time intervals. We again give an example.

Example 3.2. Consider a community of nine agents that share our earlier language $\mathcal{L}$. Let the bit string representations of their initial belief states be

1. 1100	4. 1000	7. 0000
2. 1101	5. 1101	8. 1101
3. 0001	6. 0001	9. 0001

Assume that $D = 1$. Then, for instance, the set of peers of agent 1 is initially (after 0 updates): $X_1(0) = \{1, 2, 4, 5, 8\}$. Agent 1 will update his theory to $x_1(1) = 1101$, given that all agents in $X_1(0)$ deem the first world possible, and hence $x_1(1)[1] = 1$; all but one of the peers deem the second world possible, so $x_1(1)[2] = 1$; all peers deem the third world impossible, so $x_1(1)[3] = 0$; and although x_1 initially deems the fourth world *im*possible, all other agents in $X_1(0)$ deem that world *possible*, and so $x_1(1)[4] = 1$. $\Diamond$

In Wenmackers et al. (2012), we computed the probability for an agent with a consistent belief state to arrive at an inconsistent belief state after a single update via (EHK). Except for the trivial cases with $N = 2$ or $D = 0$, we found that the probability of this event happening is always higher than zero, but lower than 2%. Moreover, we formulated some practical suggestions to avoid arriving at the inconsistent theory. For instance, it was shown that including more independent properties (increasing M) lowers the probability. Also, the members of even-numbered groups of agents (N even) have a lower probability of updating to the inconsistent theory than have the members of odd-numbered groups of comparable size. And the BCI was shown to play an important role, too: low threshold values D (narrow BCIs) result in low dynamicity, so the probability of any change in belief state is low, so a fortiori the probability of arriving at an inconsistency is low; very high bounds of confidence (D close to 2^M) were also shown to decrease the chance of updating to the inconsistent theory.

The mere possibility of arriving at an inconsistent theory—even though it has a low probability—might be thought to discredit EHK. But this would be to overlook that the update rule can have compensating advantages. The extension of the HK model that was studied in Riegler and Douven (2009) was in that paper shown to offer a clear advantage over "individualistic" updating in cases where the agents received evidence that is to some extent noisy (as evidence typically is); in such cases, the social updating led agents to approach the true theory more closely in a shorter time span. That already the simpler update rule (EHK) may offer advantages can be seen by considering agent number 7 in Example 3.2. This agent initially holds the inconsistent theory but after updating comes to hold a consistent theory. (One easily verifies that $X_7(0) = \{3, 4, 6, 7, 9\}$ and that averaging-and-rounding over the corresponding belief states results in a consistent belief state, to wit, $x_7(1) = 0001$.) However, to give a more systematic answer to the question of which advantages updating via (EHK) may have, more must be known about the properties of this update rule.

To take further steps toward determining which properties (EHK) has, beyond the ones presented in Wenmackers et al. (2012), the remainder of this paper considers this update rule again as used by a group of N agents whose belief states are theories of the world concerning M binary properties. However, now we focus our attention on the process of updating via (EHK) *repeatedly*. We achieve this by investigating the structure of the "belief space" as a whole. Due to the update rule, and starting out from a particular belief state (or theory of the world), some belief states can be reached in a single step, whereas other belief states can only be reached via intermediate steps, or cannot be reached at all. So, perhaps a larger portion of the agents will reach the inconsistent theory after repeated updating. On the other hand, agents that start out from the inconsistent theory may leave it afterwards (as just seen). *A priori*, it is not clear whether the probability of reaching the inconsistent theory after a single time step is an under- or an overestimation of the probability of reaching the inconsistent theory in general. It is good to keep in mind that, ultimately, we are not interested in estimating this probability for the model *per se*. Rather, we aim to identify useful parameters

to lower the probability of arriving at inconsistencies in actual human thinking, or to escape them once they have occurred.

Our investigations in the following focus on the case in which there is only one binary property that the agents consider to form their theory about the world (i.e., $M = 1$). In this case, there is one proposition, which can be true or false, so there are two possible worlds. There are four theories: 00 (the inconsistent theory), 01, 10, and 11 (the tautology). The Hamming distance between two different theories is either 1 (between 00 and 01, between 00 and 10, between 01 and 11, and between 10 and 11) or 2 (between 00 and 11 and between 01 and 10). It may be argued that studying $M = 1$ defeats the original purpose of modeling agents that hold theories. After all, we introduced theories of the world as a means to study agents with rich belief sets. If there is only one binary property of interest to the agents, it seems overly complicated to consider theories. Nevertheless, $M = 1$ is an important case from the theoretical viewpoint, because the relevant dynamics can be represented in three dimensions, whereas higher values of M correspond to higher-dimensional spaces, which makes it harder to visualize them. Moreover, some of the conclusions that can be reached for the $M = 1$ toy model do generalize to the higher-dimensional case. We give a brief, qualitative discussion of the general case at the end of this article.

3.1. KEY POINTS

We model a group of N agents. Their opinions concern M binary properties of the world. There are $t_M = 2^M$ possible worlds (or combinations of the properties being true or false in the world). Each agent holds a theory about the world, which can be represented as a string of t_M bits, where zero means that the agent has ruled out the corresponding possible world. There are 2^{t_M} such theories. Agents consider as epistemic peers those agents who currently hold a "sufficiently similar" theory, which means that the number of bits that are different between the agent's own theory and that of a potential peer is less than a certain threshold, called the bound of confidence D. Agents adjust their theory by averaging over the theories held by their peers. We study the resulting opinion dynamics.

4. OPINION-PROFILE SPACE

Our goal is to investigate how the opinions in the population as a whole change over time due to the iterated application of (EHK) by the individual agents. To achieve this, we first need to identify the relevant belief space, by which we mean the phase space in which we can represent the opinion dynamics of the entire group of agents. An *opinion profile* is a vector $\vec{n}$ that specifies how many agents in the entire population occupy each of the belief states (at a given point in time). In general, $\vec{n}$ has t_M components, which sum to N. (Unlike Example 3.2, the opinion profile is anonymous, so it does not keep track of which agent holds which theory.) The relevant belief space is what we will call the "opinion profile space" (OPS), in which each point represents a possible opinion profile. For $M = 1$, opinion profiles have four components, $\langle n_{00}, n_{01}, n_{10}, n_{11}\rangle$, which can be represented in a three-dimensional tetrahedron. For a representation of the tetrahedral OPS with two ($N = 2$) or three agents ($N = 3$), see **Figure 1**.

To elaborate, if there are two agents ($N = 2$), then there are ten different opinion profiles. In other words, the OPS consists of ten points, which are shown at the left-hand side of **Figure 1**. Four of these ten opinion profiles represent a consensus: $\langle 0, 0, 0, 2\rangle$, in which the two agents agree on theory 11; $\langle 0, 0, 2, 0\rangle$, in which the two agents agree on theory 10; $\langle 0, 2, 0, 0\rangle$, in which the two agents agree on theory 01; and $\langle 2, 0, 0, 0\rangle$, in which the two agents agree on theory 00. The remaining six points in the OPS represent opinion profiles in which each agent holds a different position: $\langle 0, 0, 1, 1\rangle$, in which one agents holds theory 11 and the other holds 10; $\langle 0, 1, 0, 1\rangle$, in which one agents holds theory 11 and the other holds 01; and so on. Thus, in the case with $M = 1$ and $N = 2$, the only points that can be occupied in the OPS are the four vertices of a tetrahedron (consensus) and the six midpoints of the edges (disagreement).

If there are three agents ($N = 3$), then there are twenty different opinion profiles, corresponding to an OPS that consists of twenty points, as can be seen on the right-hand side of **Figure 1**. There are still four possible opinion profiles that represent a consensus—$\langle 0, 0, 0, 3\rangle$, $\langle 0, 0, 3, 0\rangle$, $\langle 0, 3, 0, 0\rangle$, and $\langle 3, 0, 0, 0\rangle$—corresponding to the vertices of the tetrahedral OPS. There are

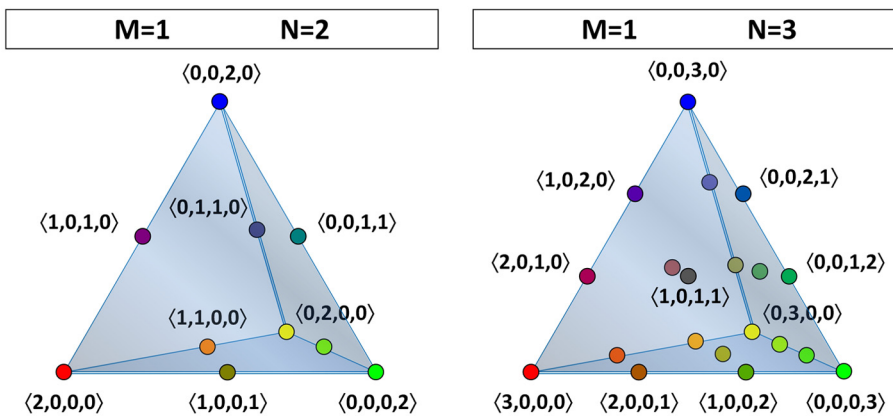


FIGURE 1 | The belief space or opinion-profile space (OPS) for the theories in a language with one atomic proposition ($M = 1$) can be visualized as a discrete grid in a tetrahedral volume. The number of grid points depends on the population size. The OPS of a population of two agents is shown at the left ($N = 2$), while that of three agents is shown at the right ($N = 3$). The grid points are indicated by colored dots and are labeled by their opinion-profile coordinates between angle brackets. (For $N = 3$, only the four vertices and the opinion profiles located in the front face are labeled.)

twelve profiles in which two agents agree and the third one does not: two on each of the six edges in the OPS. And there are four ways in which all of the agents can disagree with each other; these opinion profiles each correspond to a point on one of the four faces of the OPS.

For any fixed number of N agents (and some number M of propositions) the opinion profile space is discrete and contains $\frac{(N+t_M-1)!}{N!(t_M-1)!}$ points (this is the (hyper-)tetrahedral number of order $N+1$ in t_M-1 dimensions, or the multiset coefficient of choosing N times with repetition out of t_M options). If there are four or more agents, then the points in the OPS also occupy the interior volume of the tetrahedron. (For four agents, this concerns only the central point $\langle 1, 1, 1, 1\rangle$).

In principle, the OPS for any particular N can be computed by hand: for each possible opinion profile, one can determine each agent's peer group and apply the two-step update rule. In practice, however, a computer is required to assist in these computations, since the aforementioned number of opinion profiles in the OPS grows rapidly with N. To this end, we have written a program in Object Pascal. Instead of iterating the process for each opinion profile until it reaches a fixed point, we instructed the program to link up opinion profiles that reach a fixed point, via intermediate opinion profiles. In section 5, we will show how to abstract from the number of agents in the population (by looking at the opinion density instead of the opinion profile), but first we introduce the dynamics on the OPS brought about by social updating via (EHK).

4.1. RESULTS: DYNAMICS ON THE OPINION PROFILE SPACE

We view the OPS equipped with the two-step social update rule (EHK) (with N agents and a threshold value D) as a discrete dynamical system. Even before we look at the results, we can give a qualitative description of the dynamics. For any value of D, certain opinion profiles will act as fixed points. Populations that start out with an opinion profile outside a fixed point may be driven either toward a certain fixed point ("sink," or stable equilibrium, or attractor) or away from it ("source" or unstable equilibrium). All unstable points that are attracted toward a particular sink belong to the "basin" of this sink.

The lower the threshold D, the more fixed points we expect to find in the OPS. In the case with $D = 0$, there is no dynamics at all: the agents do not take into account any other opinions, so there is no process of social updating, and all the points in the OPS act as fixed points. (Since there is no dynamics, we cannot classify the points as sources or sinks; rather, this is a case of indifferent equilibrium.) As the BCI increases, an growing number of other opinions may be taken into account and fewer opinion profiles are fixed points. When the BCI is maximal (i.e., $D = t_M$), the dominant sources and sinks are revealed. Opinion profiles in which the agents all agree on the same theory are sinks.

We will represent the dynamics on the OPS by arrows that point from an initial opinion profile toward the corresponding final state. For the sake of illustration, we consider populations in which none of the agents hold theory 01, so that we can limit ourselves to one face of the OPS. First, suppose that there are just two agents. In this case, there is no dynamics, irrespective of the value for D. (After all, when the average is exactly equal to 1/2, the corresponding bit keeps its initial value. Hence, an agent can never be swayed by a single peer and vice versa.) This situation is illustrated in **Figure 2**: all the opinion profiles are fixed points, so there are no arrows connecting any of them.

If there are three agents, then for $D = 1$ and $D = 2$ there is some dynamics: **Figure 3** shows us that the consensus positions (at the vertices) act as sinks. For $D = 1$, there is a certain asymmetry in the face of the OPS that we are considering: there are two opinion profiles that move toward the consensus position at $\langle 0, 3, 0, 0\rangle$, but only one opinion profile each that moves toward the consensus positions at $\langle 3, 0, 0, 0\rangle$ and $\langle 0, 0, 0, 3\rangle$. To understand why not all directions in the tetrahedron are equivalent, we have to remember that there are two pairs of theories that have a larger Hamming distance between them than the other six pairs, one pair being 00 and 11, the other pair being 01 and 10. Therefore, also the two edges connecting opinion profiles corresponding to a consensus on such a pair of "more distant" theories are qualitatively different from the other six edges. In **Figure 4**, the two edges connecting consensus on "more distant" theories are indicated by a double line, whereas the four other edges are represented by a single line. Since each face has two "single" edges and one "double" edge, the analysis of each of the four faces is equivalent. The right-hand side of **Figure 4** also illustrates that the four vertices are equivalent in the sense that they all attract two other opinion profiles (for $D = 1$). The asymmetry between the edges of a single face that appeared for $D = 1$ is absent for $D = 2$, where each sink attracts two other points (at least on the face that we are considering; it attracts three points in total). The explanation for this restoration of symmetry is that, with the maximal value for

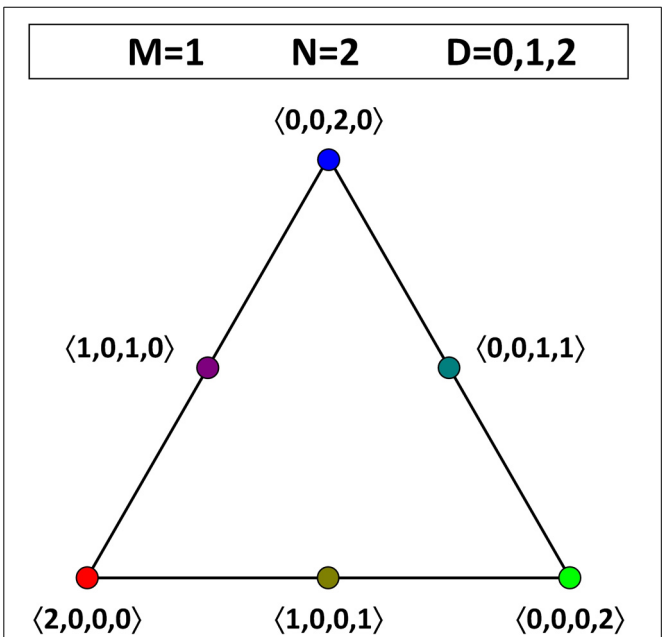


FIGURE 2 | Opinion dynamics for theories in a language with one atomic proposition ($M = 1$) and a population consisting of two agents ($N = 2$). One face of the OPS is shown. Irrespective of the value of the threshold D (D is equal to 0, 1, or 2), all opinion profiles are fixed points (including the ones not shown).

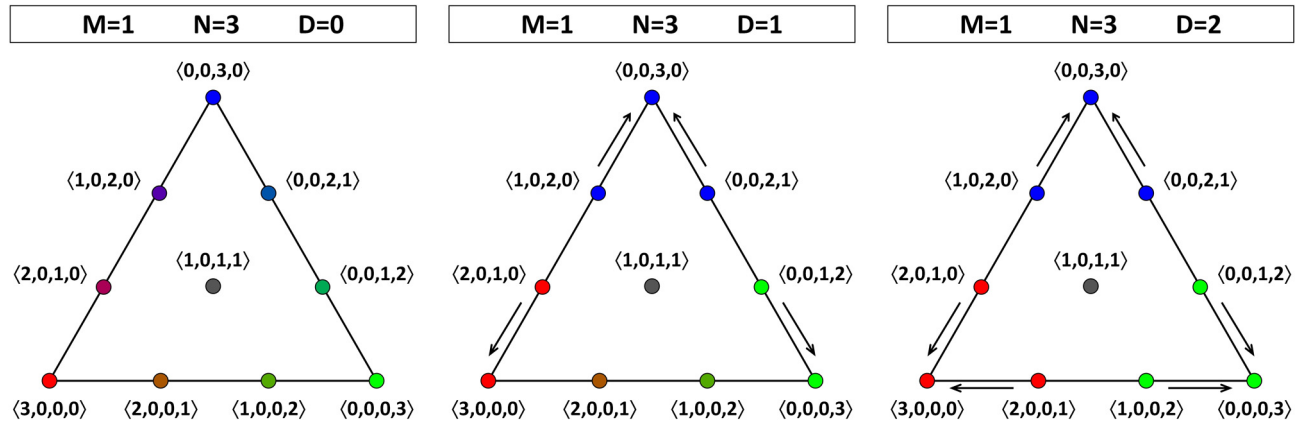


FIGURE 3 | Opinion dynamics for theories in a language with one atomic proposition ($M = 1$) and a population consisting of three agents ($N = 3$). One face of the OPS is shown. For $D = 0$ (left), there is no dynamics; all opinion profiles are fixed points. For $D = 1$ (middle), the three vertices act as sinks, but the points on the lower "double" edge (connecting profiles with a consensus on "more distant" theories—see the main text for details) show a different dynamics than those on the two "single" edges. For $D = 2$ (right), the three vertices act as sinks, and the asymmetry between points on the edges is removed.

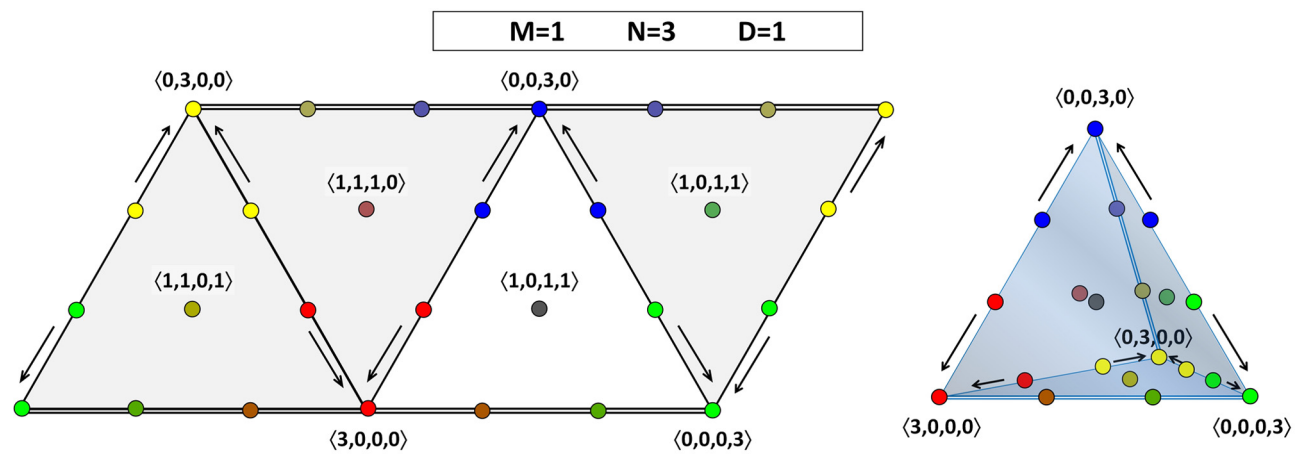


FIGURE 4 | Opinion dynamics for theories in a language with one atomic proposition ($M = 1$) and a population consisting of three agents ($N = 3$) for $D = 1$. On the left-hand side, all the faces of the tetrahedral OPS are shown side by side. The face from **Figure 3** is shown in white, the others in gray. Observe that the outer edges and their vertices are shown multiply: the leftmost diagonal edge is equal to the rightmost diagonal edge, and the left horizontal edge at the top (bottom) is the mirror image of to right horizontal edge at the top (bottom). On the right-hand side, the tetrahedral OPS is shown. This three-dimensional view allows us to verify that each vertex attracts two other opinion profiles.

D, even agents that hold maximally different theories regard each other as peers. So, unlike for $D < 2$, they do influence each other in updating their belief states. The point at the middle of the face, $\langle 1, 0, 1, 1\rangle$, is a non-attracting fixed point (source).

Figure 5 shows the opinion dynamics for a population of four agents. Although there are more points in this OPS, the results are comparable to those for $N = 3$: there is no dynamics for $D = 0$, and there is an asymmetry for $D = 1$ that is absent for $D = 2$. The three vertices are sinks and each of the three points at the middle of an edge is a source.

So far, we have considered the opinion dynamics for a fixed number of agents in the population. If we continue the above analysis for ever larger population sizes, predictable patterns appear, such as (for $D > 0$):

- The vertices act as sinks, but the number of points attracted to them depends on the BCI.
- If the number of agents is even, the midpoint of the edges is accessible and acts as a source (for other points on the edge).
- If the number of agents is a multiple of three, the midpoint of each of the faces is accessible and acts as a source (for $D = 1$).
- If the number of agents is a multiple of four, the midpoint of the entire tetrahedron is accessible and acts as a source.

This suggests a different way of studying the opinion dynamics: instead of considering populations with a particular population size, one can consider populations in general and ask, for each possible opinion, which *fraction* of a population holds the opinion (this will be called the *density* the opinion has in the population).

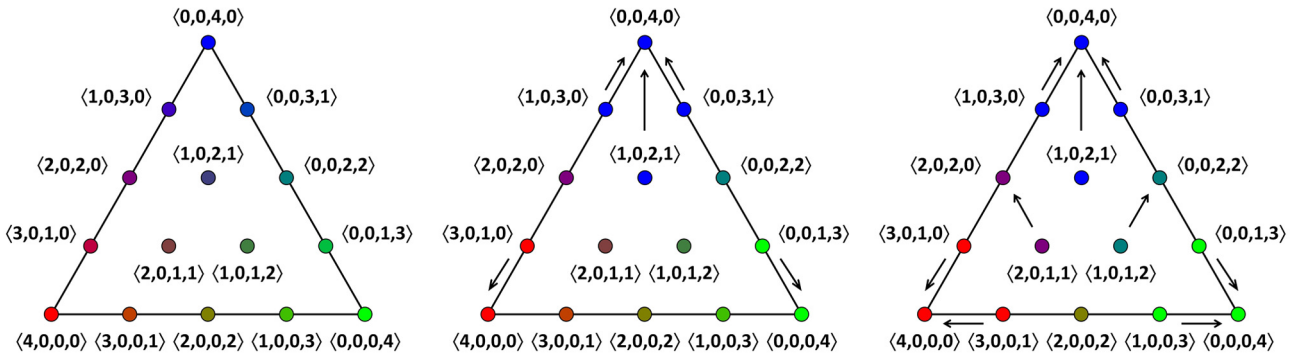


FIGURE 5 | Opinion dynamics for theories in a language with one atomic proposition ($M = 1$) and a population consisting of four agents ($N = 4$). We show the same face of the OPS for $D = 0$ (left), $D = 1$ (middle), and $D = 2$ (right).

This in turn allows one to derive the above rules immediately, without the need for considering a large number of different population sizes. With the density-based information, one can still draw conclusions for particular population sizes. For instance, if 100% of the agents hold the same opinion, that represents a consensus (point at a vertex), which can occur in populations with any number of agents ($N = 1, 2, 3, \ldots$). And if 50% of the agents hold one theory and 50% hold the other theory, there is a tie between two theories (midpoint of an edge); this can occur only in even-numbered populations ($N = 2, 4, 6, \ldots$). In the next section, we consider such opinion densities. But first, we give a probabilistic interpretation concerning the results of the dynamics on the OPS.

4.2. PROBABILISTIC INTERPRETATION OF THE OPS

We can give a probabilistic interpretation of the previous results. For instance, we may be interested in the probability that the agents in the population reach a consensus on a particular theory. If we assume that each initial opinion profile is equally likely (uniform prior probability), then the probability of reaching consensus on a particular theory is equal to the number of opinion profiles in the basin of this consensus position divided by the total number of points in the OPS. (For a non-uniform prior probability, we may compute a similar fraction based on weighted sums).

For $M = 1$ and $N = 2$, there are 10 points in the OPS and there is no dynamics, so the only way the population can end up in a consensus is by already starting out from that opinion profile. Hence, the probability of reaching consensus on a particular theory is 1/10. (The total probability of reaching a consensus is 4/10.) For $N = 3$, there are 20 points in the OPS. For $D = 0$, there is no dynamics, so the probability of reaching consensus on a particular theory is 1/20. (The total probability of reaching a consensus is 4/20, or 1/5.) For $D = 1$, two additional opinion profiles evolve toward each consensus position, so each basin consists of three points and the probability of reaching consensus on a particular theory is 3/20. (The total probability of reaching a consensus is 12/20, or 3/5.) For $D = 2$, each basin consists of four points and the probability of reaching consensus on a particular theory is 4/20, or 1/5. (The total probability of reaching a consensus is 16/20, or 4/5.) Given the nature of our update rule (EHK), it is not surprising that we find larger BCIs (larger values for D) to correspond with higher probabilities of reaching a consensus.

In our previous paper (Wenmackers et al., 2012), we only considered the probability that an agent, who starts from a consistent theory, updates to the inconsistent theory. For $M = 1$, this probability is zero. In general, there are $\frac{(N+t_M-2)!}{N!(t_M-2)!}$ opinion profiles in which no agent adheres to the inconsistent theory (i.e., the (hyper-)tetrahedral number of order $N + 1$ in $t_M - 2$ dimensions, or the multiset coefficient of choosing N times with repetition out of $t_M - 1$ options). For $M = 1$, these inconsistency-free opinion profiles are represented on a single face of the tetrahedral OPS—the face which has as its vertices each consensus on one of three consistent theories—and none of these evolve to consensus on the inconsistent theory. To investigate the phenomenon of consistent-to-inconsistent updating, we have to consider cases with larger values of M, as we did in our previous study (in which we assumed a uniform prior, not over all anonymous opinion profiles, but over the non-anonymous opinion profiles in which no agent adheres to the inconsistent theory).

4.3. KEY POINTS

An (anonymous) opinion profile specifies the number of agents that holds each of the theories. So, an opinion profile consists of 2^{t_M} numbers that add up to N, the total number of agents in the population. We consider the space of all possible opinion profiles, the OPS. The dynamics on this space shows the group-level or aggregate effect of the individual updating by the rule introduced in the previous section. Some opinion profiles act as fixed points: once the population reaches such a state, there is no further dynamics. Consensus positions are stable fixed points, which "attract" nearby opinion profiles; equally balanced (or polarized) opinion profiles are unstable fixed points, which "push away" nearby opinion profiles. By counting states in the OPS and assigning priors probabilities to initial opinion profiles, we can give a

probabilistic interpretation to the results. The analysis in terms of an OPS requires the choice of a particular population size, N; in the next section, we follow a slightly different approach that does not require this.

5. OPINION DENSITY SPACE

To simplify the analysis, we leave the number N of agents open and represent all possible opinion profiles (for arbitrary N) simultaneously, using the opinion density space (ODS). For a given opinion profile $\overrightarrow{n}$, the corresponding opinion density $\overrightarrow{d}$ can be found via normalization, that is, division by the number N of agents: $\overrightarrow{d} = \overrightarrow{n}/N$. Like $\overrightarrow{n}$, $\overrightarrow{d}$ is a vector with t_M components. We represented the components of a particular opinion profile $\overrightarrow{n}$ between angle brackets, $\langle \ldots \rangle$; although confusion is unlikely, we will represent the components of an opinion density $\overrightarrow{d}$ between round brackets, $(\ldots)$. The opinion density coordinates can be viewed as barycentric coordinates, specifying which *fraction* of the agents adheres to each theory[6].

Another way of looking at the transition from OPS to ODS is as follows: we can track the dynamics for a large set of different population sizes and represent the accumulated data in a single tetrahedral grid. In the limit where we combine the OPSs for all (infinitely many) finite population sizes, this accumulative OPS becomes continuous instead of a discrete grid. Hence, the ODS is a continuous space in $t_M - 1$ dimensions. (There are t_M components of the opinion density vector, which are fractions that sum to 1, so there remain $t_M - 1$ degrees of freedom).

To visualize the ODS, we have written an additional program in Object Pascal. Although the ODS represents a continuous space, numerical methods require it to be discretized, such that the program only encounters density vectors which have four rational indices. By multiplying the four rational indices of an opinion profile by their least common denominator, we compute an opinion profile that is representative of that density. The evolution of this profile is computed as before. The numerical result is indicated by means of colors (as explained below).

[6]The notion of barycentric coordinates, which comes from geometry, may need some introduction. We first introduce the concept of a simplex. A two-dimensional simplex is a triangle: a figure with three vertices. In three dimensions, a simplex is a tetrahedron: a figure with four vertices. In general, in k dimensions, a simplex is a figure with $k+1$ vertices. The ODS is a simplex with $2^{t_M} = 4$ vertices. Hence, for the simplest case with $M = 1$, we are dealing with a tetrahedral ODS in $k = 2^{t_M} - 1 = 3$ dimensions. To indicate a particular point inside a k-dimensional simplex, one can use k Euclidean coordinates (belonging to k orthogonal axes), but for many applications it is more natural to use barycentric coordinates. The word "barycenter" refers to the center of mass, and barycentric coordinates indicate how much a point "gravitates" toward each of the vertices of the simplex. Since a k dimensional simplex has $k+1$ vertices, it also has $k+1$ barycentric coordinates, but since those coordinates are fractions that sum to unity, there are only k degrees of freedom. In three dimensions, a barycentric plot indicates the ratios of four quantities. The geometric center of a k-dimensional simplex is characterized by $k+1$ barycentric coordinates that are all equal to $1/(k+1)$. In general, points inside the (hyper-)volume of the simplex have barycentric coordinates that are all strictly positive. Points with one barycentric coordinate equal to unity and all the others equal to zero indicate a vertex position.

5.1. RESULTS: DYNAMICS ON THE OPINION DENSITY SPACE

We consider the ODS equipped with (EHK) as update rule (for particular values of D) as a continuous dynamical system.

As before, we focus on the case with $M = 1$. In this case, opinion densities have four digits, which are fractions that sum to 1, so there remain three degrees of freedom. Hence, these opinion densities can be represented using barycentric coordinates in a three-dimensional tetrahedron (inside the volume as well as on the surface). At the four vertices of the tetrahedral ODS for $M = 1$, we find the opinion profiles that have all their weight concentrated on a single theory, corresponding to populations in which all the agents agree on the same theory (consensus). On the edges of the tetrahedron, we find populations in which only two of the four theories are represented (the other two having density zero). On the faces of the tetrahedron, we find populations in which one of the four theories is not represented. Inside the volume of the tetrahedron, in each population there is at least one agent for each theory, so none of the density components is zero.

Also similar as before, we only represent a single face of the tetrahedral ODS: the triangle with vertices at $(0, 0, 0, 1)$, $(0, 0, 1, 0)$, and $(1, 0, 0, 0)$, with the "double" edge at the bottom. Within this triangle, all opinion profiles have zero density at the second position: there are no agents that hold the theory 01.

For each position in the chosen triangle, we compute the (normalized) opinion profile that it will ultimately evolve to. We represent this by a color. Specifically, the color (R, G, B) (with $R, G, B \in \{0, \ldots, 255\}$) indicates that the opinion profile at that position will evolve to the opinion profile with barycentric coordinates equal to $(G/255, 0, B/255, R/255)$. For instance, the redder a point, the larger the fraction of agents that will finally adhere to the inconsistent theory, 00. The results depend on the threshold value D and are presented at the left-hand side of **Figure 6**. For each point, we also indicate after how many steps the final state is reached. We represent this with a gray-scale on the right-hand side of **Figure 6**.

The results for $D = 0$ are trivial: the agents do not take the opinions of others into account, so there is no dynamics. On the right-hand side of **Figure 6**, we see that all the positions have the color corresponding to the initial opinion profile. At the left-hand side of **Figure 6**, we see that zero steps are required to reach the final state. Both observations confirm that all opinion profiles are fixed points. Because there is no dynamics, it is a situation of indifferent equilibrium. (This image is still helpful, because—due to the absence of dynamics—each point in it is colored based on its *own* coordinates, which can be used as a key to interpret the representation of the results with dynamics.)

The results for $D = 2$, the maximal threshold value in the case of $M = 1$, do show dynamics. In the colored image, we see clear evidence that a "double" edge of the tetrahedron was positioned at the bottom: it leads to a bilateral symmetry of the pattern. There are six fixed points. The three consensus positions at the vertices are fixed points, which act as sinks for large portions of the face. The three positions halfway along the edges are fixed points as well. Those on the "single" edges each attract opinion profiles from a line in the triangle; the fixed point on the "double" edge acts as a sink. The gray-scale image confirms these findings: the

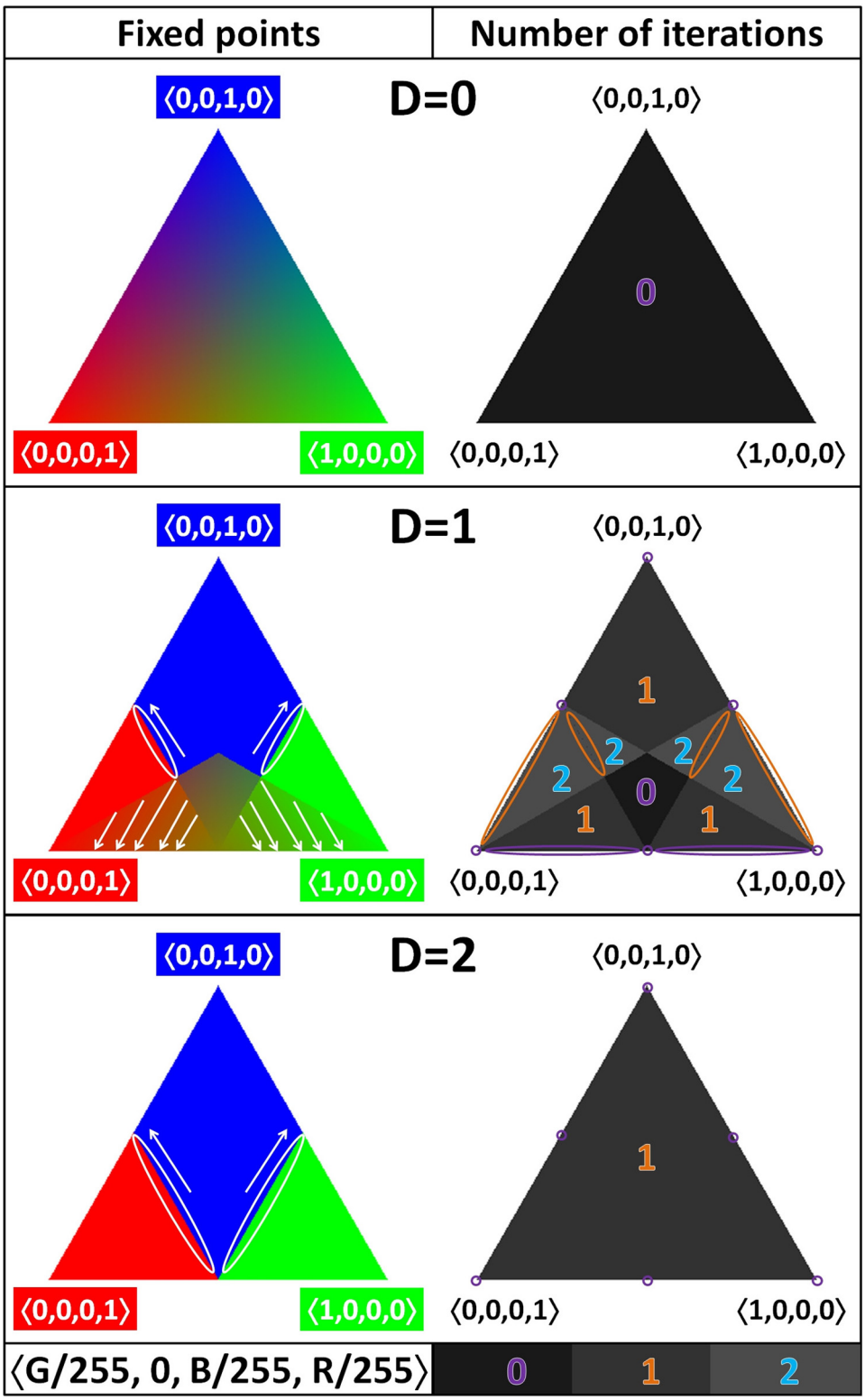


FIGURE 6 | Opinion dynamics for $M = 1$: one face of the opinion density space (ODS). On the left-hand side, each position in the ODS is colored depending on its final state (see main text for details); smaller features are indicated with white ellipses and arrows. On the right-hand side, each position in the ODS is given a gray-scale value depending on how many rounds of updating are required for it to reach its final state; smaller features are indicated with ellipses (purple for zero, orange for one, and blue for two). The value of the BCI threshold is varied: top row $D = 0$ (minimal), middle row $D = 1$ (intermediate), and bottom row $D = 2$ (maximal).

six fixed points do not require any iterations, whereas the others settle after just one update.

Intermediate values for D tend to lead to more complex and interesting behavior. This general trend holds up even for $M = 1$, although there is only one intermediate value: $D = 1$. The bilateral symmetry (and lack of additional symmetry), already observed for $D = 2$, is present here, too, but both the color and the gray-scale image show further features. There are fewer fixed points than for $D = 0$, but more than for $D = 2$: there is a kite-shaped region of fixed points (indifferent equilibrium), and the "double" edge consists of fixed points, all of which act as sinks for a line in the triangle. Moreover, this is the only case with $M = 1$ for which some initial opinion profiles require two rounds of updating to arrive at the final state.

Recall that for a fixed number of agents, not all points of the continuous ODS are accessible. Once you have computed the opinion dynamics for the ODS, you can use the results to construct the dynamics on an OPS for a fixed number of agents, N, by locating a density that is accessible for the N of interest and using the color of that point to determine to which opinion profile it will evolve. (In fact, the results on OPSs in the previous figures do already use the same color convention as that used for the ODS).

5.2. PROBABILISTIC INTERPRETATION OF THE ODS

Similarly to the discussion of the OPS results, we also give a probabilistic interpretation of the results concerning the ODS. If we assume that each initial opinion profile is equally likely (uniform prior probability), then the probability of reaching consensus on a particular theory is equal to the volume of the basin associated with this consensus position divided by the total volume of the OPS. At least, this fraction expresses the limit probability associated with an infinite population size, in which the relative importance of special points (unstable equilibria) is vanishingly small.

In **Figure 7**, we illustrate the four basins associated with the four consensus positions in the ODS of $M = 1$ and $D = 2$. Each basin has the same shape with five faces: two equilateral triangles and one rhombus that face the exterior of the ODS and two isosceles right triangles that face the interior of the ODS (see also Supplementary Material). The four basins have one common edge (at the interior, where the isosceles right triangles meet) that connects the midpoints of the two "double edges" of the tetrahedral ODS.

Since the four basins have the same shape and size and together fill the entire volume of the ODS, they each correspond to a relative volume of 1/4. Under the assumption of a uniform prior, the limit of the probability of arriving at a particular consensus for exceedingly large populations is 1/4 ($M = 1$ and $D = 2$). For maximal D, the limit probability of arriving at some consensus is 1. Under these conditions, the unstable equilibria on the edges of the basins are isolated points, lines, or areas, which have zero volume and thus zero probability.

In particular, in the infinite population limit there is a probability of 1/4 of arriving at the inconsistent theory. However, if we only consider opinion densities where the inconsistent theory initially has zero density (which are all represented at the a single face of the ODS), the probability of evolving to an opinion profile with a non-zero density at the inconsistent theory (let alone unit density at this position) is zero (at least for $M = 1$).

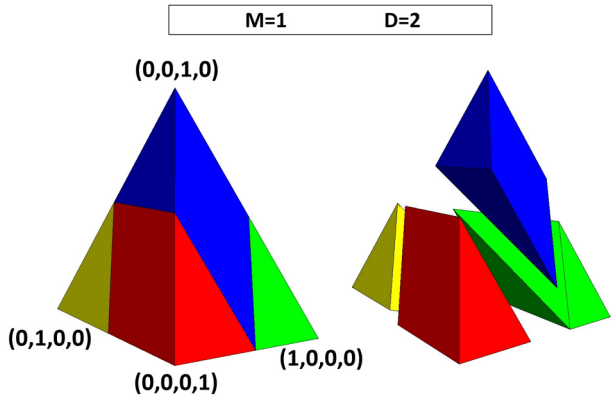


FIGURE 7 | Shape of the basins in the opinion density space (ODS) for $M = 1$ and $D = 2$. Left-hand side: three-dimensional view of ODS with the face shown in previous figures turned toward the right. Right-hand side: exploded view of the same ODS, showing the four basins separately. Each basin has the same shape (with five faces: one rhombus, two equilateral triangles, and—facing the interior of the ODS—two isosceles right triangles) and a volume that occupies one quarter of the tetrahedral ODS.

5.3. KEY POINTS

Whereas the discrete OPS depends on a particular population size, N, the continuous ODS represents the density of theories in populations of arbitrary size. By considering volumes in the ODS and assigning a prior probability distribution to initial opinion profiles, we can give a probabilistic interpretation to the results, which serve as a good approximation for very large population sizes, but does not apply to small groups. We observe that even if special points (such as stable fixed points) make up a small portion of the ODS, these points tend to be represented in small populations (causing the dynamics to end after few rounds of updating).

6. GENERAL DISCUSSION

Due to social updating, an agent who starts out with a consistent theory about the world may arrive at the inconsistent theory. Even if maintaining consistency at all times is too demanding for non-ideal beings to qualify as a necessary condition for rationality (Cherniak, 1986), it is presumably something that rational beings should aim for. This may suggest that social updating is a vice, from the perspective of rationality. However, in our first study (Wenmackers et al., 2012) we computed the probability for an agent to update to the inconsistent theory and found it to be non-zero, but relatively small (lower than 2%); moreover, it can be made arbitrarily low by strategically varying the model parameters.

Our current study of the opinion dynamics on the belief space reveals another virtue of the social updating process: even if an agent starts out at the inconsistent theory, the agent's opinion may change—to one of the consistent theories—due to the social update rule. This could already be seen on the basis of

Example 3.2, but the results depicted in **Figure 6** give more systematic information in this respect: except for the rightmost edge and its two vertices, all the opinion profiles in the presented face of the tetrahedron contain at least one agent who starts out at the inconsistent theory. Nevertheless, when there is any dynamics at all, many of these opinion profiles evolve to different profiles, some of which have no agents at the inconsistent theory. This is true, in particular, for all the opinion profiles in the blue and green areas, which act as basins for consensus positions on consistent theories.

We have given a probabilistic interpretation of the results on the belief space (OPS and ODS). We have seen that in the limit for an infinite population size and for large BCIs ($D = w_M$), the relative importance of unstable equilibria vanishes. For $M = 1$ and $D = 2$, the probability of arriving at a population-wide consensus on some theory is unity. In particular, the probability of arriving at a population-wide consensus on the inconsistent theory is 1/4. Once the agents reach consensus on the inconsistent theory, there will be no further dynamics, because all consensus positions are fixed points. Hence, this result may be regarded as a worst case. However, this case study is highly unrealistic for (at least) three reasons.

First, the assumption of a uniform prior on the opinion profiles does not apply to real cases. Observe that if the agents were to pick out their initial theory at random, the distribution of initial anonymous opinion profiles would be higher around the center of the belief space. (For larger populations, there are more combinations of individual theories that lead to an anonymous opinion profile, in which all theories are represented almost evenly.) More importantly, however, we do *not* expect the agents to adopt an initial theory at random but rather to possess some prior knowledge, such that the distribution of their initial theories is clustered around the true theory (which is necessarily a consistent one). Hence we also expect a preferential position of the opinion profiles in a region around consensus on the true theory. For this reason, investigation of a more complex model, based on a variant of our current update rule (EHK), but including evidence-gathering as well as social updating, is high on our to-do list.

Second, in many practical situations relevant population sizes tend to be small (just think of the last meeting you attended), such that the infinite population limit does not apply well to them. In smaller populations, the relative importance of unstable equilibria (which do not lead to consensus) is more pronounced.

Third, modeling belief states as theories of the world only has practical relevance when $M > 1$, for which the relative size of the basins associated with consensus positions decreases rapidly (as $1/t_M$).

For all these reasons, we estimate the probability of arriving at a consensus on the inconsistent theory to be very small in a realistic setting—in any case well below 1/4.

The mechanism for social updating may also be criticized in the following way. If agents' belief states are theories, their beliefs are closed under the consequence relation. So, illustrating with theories for the case of $M = 1$ (cf. Example 3.2), an agent whose belief state is characterized by the string 1100 is supposed to believe also the propositions coded as 1110, 1101, and 1111. This is not reflected in our current update rule (EHK) and suggests an asymmetric composition of the peer group: for $M = 1$ and $D = 1$, an agent A with theory 1111 and an agent B with theory 1100 are not each other's peers according to our current model. However, agent B also ought to believe A's theory, but not vice versa. We may now suggest an alternative way of determining an agent's peer group: by taking into account also those agents that hold a theory which is within distance D of at least one of the consequences of the first agent's theory. Doing so would help to protect agents against updating to the inconsistent theory. However, it also introduces a preference for less informative theories, so it may hamper the agents' chances of finding the (strongest) true theory. Hence, this is a case where different epistemic goals (rationality versus finding the truth) are in direct conflict with each other and selecting the optimal normative model seems to require meta-norms of rationality.

In our previous work (Wenmackers et al., 2012), we have considered the probability of arriving at an opinion profile in which at least one agent adheres to the inconsistent theory, starting out from an opinion profile without any such agent (and assuming a uniform prior over these anonymous profiles). We found this probability to be zero for $M = 1$. This finding is confirmed in the current study. Nevertheless, by studying the dynamical space in general, we have observed certain trends that help to explain the previously obtained results for the probability of consistent-to-inconsistent updating.

For $M = 2$, the probability that an agent will arrive at the inconsistent theory, in a population where none have adopted this theory, is non-zero (provided that $D > 0$ and $N > 2$). In our previous work, we observed that this probability decreases when more independent issues are considered (that is, when M increases beyond 2). We are now in a better position to explain the—essentially combinatorial—mechanism behind this finding. Although we have not presented cross-sections for the higher-dimensional case, we can give a qualitative discussion of cases with $M > 1$. As M increases, the belief space becomes higher-dimensional ($t_M - 1$) and the basin that is attracted by the sink corresponding to consensus on the inconsistency becomes a smaller fraction of its total (hyper-)volume (equal to $1/t_M$ for $D = w_M$). This corresponds to the observation in our previous study that the probability of updating to the inconsistent theory is lowered by forming theories over more independent issues (higher M). For a larger number of agents (higher N), the dimensions of the belief space remain the same, but the opinion profile has access to more points of this space. As a result, the probability of consensus on the inconsistent theory is lower, too; this is in line with the earlier findings as well.

For belief spaces with a fixed number of agents (with $M = 1$ and $D > 0$), we observed that if the number of agents is even, the midpoint of the edges is accessible and acts as a source (in respect to other points on the edge). This is confirmed by our study of the ODS: the midpoint of an edge belongs to a line separating two or three basins. In the ODS, it also becomes clear that the midpoint on a "single" edge acts as a sink for points from the line between this midpoint and the midpoint of a "double" edge (half of the line for $D = 1$, all of it for $D = 2$). Moreover, if the number of agents is a multiple of four, the midpoint of the

entire tetrahedron is accessible and acts as a source. In contrast, if the number of agents is a multiple of three, the midpoint of each of the faces is accessible and acts as a source (for $D = 1$). So, in the case of an even number of agents, there are more fixed points than in the case of an odd number of agents. Taken together, these effects explain the "even–odd wobble" in our previous study: the observation that agents have a lower probability of updating to the inconsistent theory in an even-numbered population than in an odd-numbered population of similar size.

Moreover, for fixed M, there is a limited number of these special points, whereas the total number of accessible points in the belief space rises fast when the number of agents, N, increases. Consequently, the number of these special points as compared to the total number of opinion profiles in the hyper-volume decreases when N increases, which explains the attenuation of the wobble for larger populations. If we consider (a face of) the ODS for $M = 1$ and $D > 0$ (cf. **Figure 6**), we see that the majority of opinion densities belong to some basin that is attracted to a sink. However, most of the points that are accessible in the OPS for a relatively small population size do not belong to these basins. Hence, small populations have a relatively high probability of producing delicately balanced opinion profiles, which tend to act as unstable equilibria (sources) and do not lead to full consensus.

Additionally, as the number M of propositions increases, the dimensionality of the belief space increases, as does the absolute number of these special points, but their number as compared to the possible points in the hyper-volume decreases. This explains the earlier observed decrease in the maximal probability of updating to the inconsistent theory as M increases.

While the model studied in this paper is idealized in several respects, it is not completely unrealistic. Even if real agents do not generally compromise with their peers exactly in the way our artificial agents do, real agents do tend to influence each other's belief states, whether consciously or not. Idealized models can give information about such processes, much in the way in which the Ideal Gas Law gives information about the behavior of real gases. Also, there are several ways to make the model more realistic, for instance, as indicated earlier, by providing the agents with direct evidence about the truth, which in our model could be added as a driving force, directed toward a particular theory, or—equivalently—as an external potential directed toward one of the vertices of the ODS, corresponding to consensus on a theory with exactly one non-zero bit.

But even in its present, idealized form, the model we have studied demonstrates that there may be issues of rationality specifically arising from the way or ways we interact epistemically with fellow inquirers. We will be content if this sways some traditional ("individualistic") epistemologists as well as some psychologists to take the social level into consideration in their studies of rationality. For the latter group, we note that already the current model suggests a number of seemingly worthwhile empirical studies, focusing on how real people influence one another's belief states, on which factors determine whether people regard someone as their peer (in the technical sense used here), and on whether whatever epistemic interactions take place in reality tend to aid the achievement of people's epistemic goals.

FUNDING

Sylvia Wenmackers's work was financially supported by a Veni-grant from the Netherlands Research Organization (NWO project 639.031.244 "Inexactness in the exact sciences").

ACKNOWLEDGMENTS

We are greatly indebted to Christopher von Bülow for very helpful comments on a previous version of this paper. We are also thankful for helpful comments by the guest editor Shira Elqayam and the two referees.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00581/abstract

REFERENCES

Andler, D. (2012). "What has collective wisdom to do with wisdom?" in *Collective Wisdom: Principles and Mechanisms*, eds H. Landemore and J. Elster (Cambridge: Cambridge University Press), 72–84. doi: 10.1017/CBO9780511846427.005

Cherniak, C. (1986). *Minimal Rationality*. Cambridge, MA: MIT Press.

Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Adv. Complex Syst.* 3, 87–98. doi: 10.1142/S0219525900000078

Dittmer, J. C. (2001). Consensus formation under bounded confidence. *Nonlinear Anal.* 7, 4615–4621. doi: 10.1016/S0362-546X(01)00574-0

Douven, I. (2009). Uniqueness revisited. *Am. Philos. Q.* 46, 347–361. Available online at: http://www.jstor.org/stable/40606911

Douven, I. (2010). Simulating peer disagreements. *Stud. Hist. Philos. Sci.* 41, 48–157. doi: 10.1016/j.shpsa.2010.03.010

Douven, I., and Cuypers, S. (2009). Fricker on testimonial justification. *Stud. Hist. Philos. Sci.* 40, 36–44. doi: 10.1016/j.shpsa.2008.12.013

Douven, I., and Riegler, A. (2010). Extending the Hegselmann–Krause model I. *Logic J. IGPL* 18, 323–335. doi: 10.1093/jigpal/jzp059

Elga, A. (2007). Reflection and disagreement. *Noûs* 41, 478–502. doi: 10.1111/j.1468-0068.2007.00656.x

Fricker, E. (1987). The epistemology of testimony. *Proc. Aristotel. Soc.* 61, 57–83.

Gaifman, H. (1986). "A theory of higher order probabilities," in *Theoretical Aspects of Reasoning about Knowledge*, ed J. Y. Halpern (Los Altos CA: Morgan and Kaufmann), 275–292.

Goldman, A. I. (1999). *Knowledge in a Social World*. Oxford: Oxford University Press. doi: 10.1093/0198238207.001.0001

Goldman, A. I. (2001). Experts: which ones should you trust? *Philos. Phenomenol. Res.* 63, 85–110. doi: 10.1111/j.1933-1592.2001.tb00093.x

Hegselmann, R., and Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *J. Artif. Soc. Soc. Simul.* 5, 1–24. Available online at: http://jasss.soc.surrey.ac.uk/5/3/2.html

Hegselmann, R., and Krause, U. (2005). Opinion dynamics driven by various ways of averaging. *Comput. Econ.* 25, 381–405. doi: 10.1007/s10614-005-6296-3

Hegselmann, R., and Krause, U. (2006). Truth and cognitive division of labor: first steps towards a computer aided social epistemology. *J. Artif. Soc. Soc. Simul.* 9. Available online at: http://jasss.soc.surrey.ac.uk/9/3/10.html

Lackey, J. (1999). Testimonial knowledge and transmission. *Philos. Q.* 49, 471–490. doi: 10.1111/1467-9213.00154

Mercier, H., and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–111. doi: 10.1017/S0140525X10000968

Riegler, A., and Douven, I. (2009). Extending the Hegselmann–Krause model III: from single beliefs to complex belief states. *Episteme* 6, 145–163. doi: 10.3366/E1742360009000616

Simon, H. A. (1955). A behavioral model of rational choice. *Q. J. Econ.* 69, 99–188. doi: 10.2307/1884852

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York, NY: Doubleday.

van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford: Oxford University Press. doi: 10.1093/0198248601.001.0001

Weisbuch, G., Deffuant, G., Amblard, F., and Nadal, J. P. (2002). Meet, discuss and segregate! *Complexity* 7, 55–63. doi: 10.1002/cplx.10031

Wenmackers, S., Vanpoucke, D. E. P., and Douven, I. (2012). Probability of inconsistencies in theory revision: a multi-agent model for updating logically interconnected beliefs under bounded confidence. *Eur. Phys. J. B* 85, 44. doi: 10.1140/epjb/e2011-20617-8

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 25 May 2014; published online: 18 June 2014.
Citation: Wenmackers S, Vanpoucke DEP and Douven I (2014) Rationality: a social-epistemology perspective. Front. Psychol. ***5****:581. doi: 10.3389/fpsyg.2014.00581*
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 30 January 2014
doi: 10.3389/fpsyg.2014.00035

The outlandish, the realistic, and the real: contextual manipulation and agent role effects in trolley problems

Natalie Gold[1]*, Briony D. Pulford[2] and Andrew M. Colman[2]

[1] *Philosophy Department, King's College London, London, UK*
[2] *School of Psychology, University of Leicester, Leicester, UK*

Edited by:
Shira Elqayam, De Montfort University, UK

Reviewed by:
Eldad Yechiam, Technion - Israel Institute of Technology, Israel
Sebastien Tassy, Assistance Publique Hopitaux de Marseille, France

***Correspondence:**
Natalie Gold, Philosophy Department, King's College London, Strand, London WC2R 2LS, UK
e-mail: natalie.gold@rocketmail.com

Hypothetical trolley problems are widely used to elicit moral intuitions, which are employed in the development of moral theory and the psychological study of moral judgments. The scenarios used are outlandish, and some philosophers and psychologists have questioned whether the judgments made in such unrealistic and unfamiliar scenarios are a reliable basis for theory-building. We present two experiments that investigate whether differences in moral judgment due to the role of the agent, previously found in a standard trolley scenario, persist when the structure of the problem is transplanted to a more familiar context. Our first experiment compares judgments in hypothetical scenarios; our second experiment operationalizes some of those scenarios in the laboratory, allowing us to observe judgments about decisions that are really being made. In the hypothetical experiment, we found that the role effect reversed in our more familiar context, both in judgments about what the actor ought to do and in judgments about the moral rightness of the action. However, in our laboratory experiment, the effects reversed back or disappeared. Among judgments of what the actor ought to do, we found the same role effect as in the standard hypothetical trolley scenario, but the effect of role on moral judgments disappeared.

Keywords: context effects, decision making, hypothetical scenarios, responsibility, trolley problems

INTRODUCTION

Psychologists and philosophers use hypothetical dilemmas to elicit moral judgments (e.g., Kamm, 1996; Greene et al., 2001; Rozyman and Baron, 2002; Cushman et al., 2006; Schaich Borg et al., 2006; Waldmann and Dieterich, 2007; Nadelhoffer and Feltz, 2008). Psychologists aim to discover the factors that influence judgments, while philosophers use their intuitions to inform moral theorizing. The scenarios are typically fairly outlandish, involving events that are unlikely to occur in everyday life, and mostly concern life and death decisions. For instance, *trolley problems* are a family of moral dilemmas devised by philosophers in order in order to investigate why it is permissible to cause a harm to one in order to save many in some circumstances but not in others (Foot, 1967; Thomson, 1976, 1985). The paradigm trolley problem is *Side-track*: there is a runaway train that threatens to kill five men on the track ahead. An agent can save the five by switching a lever that will divert the trolley onto a side-track. However, on the side-track is one man, who would be killed. The question is whether it is morally permissible for the agent to save the five and kill the one. Other trolley problems, which are often contrasted to Side-track, vary the details about how the five are saved and the one killed.

In the original version of the trolley problem suggested by Foot (1967), the agent was the driver of the trolley. Thomson changed the agent to a passenger (Thomson, 1976) and later to a bystander (Thomson, 1985). One of the reasons that she gave for the change in role is that, as the "captain of the trolley," the driver is in a special position, being "charged by the trolley company with responsibility for the safety of his passengers and anyone else who might be harmed by the trolley he drives" (Thomson, 1985, p. 1397). In contrast, the bystander at the switch "is a private person who just happens to be there" (Thomson, 1985; p. 1397). The other reason Thomson gave is that the driver, by driving a trolley into the five, would be killing them. Hence the driver faces a choice between killing five and killing one. However, the other scenarios to which the driver is being compared involve the choice between killing one and letting five die—the predicament that is faced by the passenger and the bystander.

Thomson's bystander is now the paradigm trolley problem, but versions in which a passenger can turn the train onto a side-track have also attracted some attention from philosophers (Quinn, 1989) and psychologists (Hauser et al., 2007). Being a passenger or a bystander might also affect what the agent in the scenario ought to do. Passengers are more involved in the situation than bystanders, for whom doing nothing is, arguably, just staying out of it. Specifically, we might think of *bystanders* as onlookers, who are unexpectedly given the chance to intervene and re-direct a threat, whereas *passengers* are already participants in the situation, without being one of the people who are directly affected by the threat.

Previous experiments show that people's moral judgments about turning the train in Side-track are affected by the agent's role, as a passenger or a bystander. Pulford et al. (2012) found that 84% of subjects judged that it was morally permissible for the agent to turn the train down a side-track when she was a passenger, compared to 65% (significantly fewer) when she was a bystander. The passenger scenario replicated a dilemma from Hauser et al. (2007), which elicited a higher level of agreement

that it is morally permissible to turn the train (85%) than their other scenarios, some of which were bystander scenarios—although they did not include a bystander version of Side-track.

Side-track is one of the less outlandish versions of the trolley problem. It has even been known to occur in real life (CNN U.S., 2003). However, it is hardly a familiar occurrence. Another popular version, introduced by Thomson (1985), is *Footbridge*, where the agent can save the five by pushing a large man off a footbridge in front of the train, stopping the train but killing the one. As well as imagining an unusual scenario, responding to the Footbridge dilemma involves suspending disbelief that a large person—even one sometimes described as wearing a backpack—would be solid and massive enough to stop a train. Arguably the most far-fetched trolley problem is Frances Kamm's (1996, p. 154) *Lazy Susan* case, where the five and the one are seated on opposite sides of a giant lazy Susan, which the agent can rotate in order to save the five from the train but, in doing so, puts the one in its path.

Philosophers claim to elicit "common sense intuitions" from these scenarios, which they can use in constructing moral theories (Kamm, 1989; p. 227). Those moral theories are presumably supposed to be applicable to everyday moral decisions. However, Woodward and Allman (2007) argue that reliable judgments are the result of learning processes (which may be implicit) with corrective feedback, where feedback could include the experience of others, historical situations, or learning from cases that are analogous to the situation being assessed. Highly unrealistic cases such as trolley problems do not meet this criterion, and Woodward and Allman caution against their use in moral theorizing.

There are several reasons why there may be differences in performance between unrealistic scenarios and real life. One possibility is that mental processes which are adapted to everyday environments perform poorly when tested in an unusual context. This argument is similar to Gigerenzer's external validity critique of the heuristics and biases literature (Gigerenzer et al., 1999). A second possibility is that unusual scenarios may not elicit normal strategies and thought processes. In real life, moral cognition usually operates swiftly and implicitly, and the "extreme and unfamiliar situations such as those posed by classic moral dilemmas could evoke unusual strategies and thought processes rather than those typically used for common moral judgments" (Knutson et al., 2010; p. 379). This has led some psychologists to argue that ecological validity is crucial for studying moral judgment (Moll et al., 2005).

Most dilemmas used in research on moral judgments involve the causing or preventing of deaths, which is far from most people's everyday experience. Gold et al. (2013) found that the standard pattern of intuitions was preserved in hypothetical scenarios that were analogous to Side-track and Footbridge, but where the outcomes were economic harms, such as loss of a job, income, or property damage. This suggests the possibility of investigating judgments in trolley problems that are more familiar from everyday life. It also raises the possibility of operationalizing trolley problems in the laboratory, with subjects making moral judgments about decisions that are actually being taken, whose outcomes affect the distribution of small economic harms. It is standard to use small economic incentives in behavioral economics, including in the study of games that elicit moral behaviors such as altruism, fairness, trust, cooperation, and reciprocity (e.g., Berg et al., 1995; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Andreoni and Miller, 2002; Andreoni et al., 2002; Fehr and Schmidt, 2006).

We present two experiments designed to test whether intuitions about agent role effects in trolley problems are preserved in more familiar scenarios and in real decision-making situations[1]. We took the decision structure of Side-track, where an agent has the possibility of diverting a threat to five people with the side-effect of harming one, and transplanted it to a scenario involving a game show, a context that is familiar to most people who have watched television. The harm that would befall the one and the five involved loss of money rather than loss of life. In Study 1, we used hypothetical scenarios and we compared role effects in judgments in our game show scenario to those in the standard scenario, where the decision is whether or not to turn a train. In Study 2 we operationalized the game show scenario in the laboratory, allowing us to elicit judgments in real time about a decision that was actually being taken.

STUDY 1

We conducted a between-subjects experiment, varying the agent's role in the scenario, *onlooker* vs. *participant*, and the context of the decision. In one condition, we used the standard context of the runaway train. In the others, we changed the context to that of a game show, in which the agent can save five contestants from being knocked out and losing their winnings but, as a side effect, this leads to one other contestant being knocked out. Game shows where contestants are knocked out during the course of the game, and where contestants may have to leave the show forfeiting their winnings, are a familiar staple of television.

As well as comparing the train to a game show, we manipulated the level of the loss in the game show scenarios, comparing the *large game show* scenario, where the contestants stood to lose £200,000 (more than the average price of a house in the UK), to the *small game show* scenario, where the contestants stood to lose £10. We elicited judgments about the rightness of the action and about what the agent should do, and we asked subjects about the agent's responsibility for taking the action as well as about various other factors which may be relevant to moral judgment, and about how believable they found the scenario.

METHODS

Subjects

There were 1215 subjects: 359 men, 761 women, and 95 people who did not disclose their gender. Subjects were mainly voluntary visitors to an on-line survey, which they completed in their own time, after following a link to a SurveyGizmo online data collection website. The survey was promoted online, including at http://psych.hanover.edu/research/exponnet.html, and through UK university e-mail lists. There were 31 subjects who voluntarily participated in a pen and paper version

[1]We had originally hoped to compare Side-track and Footbridge, but we struggled to come up with any real life examples of dilemmas with a similar structure to Footbridge—grist to the mill of Woodward and Allman's (2007) argument.

distributed in an undergraduate philosophy class at the University of Edinburgh. Subjects were not paid for their participation. The majority of the subjects (67%) were British or American, the rest came from all over the world; 75% spoke English as their native language. Subjects were aged between 18 and 72 years ($M = 24.87$, $SD = 8.83$).

Materials

We compared six scenarios in a 2 (Role: *onlooker* vs. *participant*) × 3 (Context: *train* vs. *large game* vs. *small game*) experimental design. The *train* scenarios were based on the standard trolley problem where the agent has the possibility of turning a train onto a side-track, saving five lives at the cost of one. We varied whether the agent was a bystander on the tracks (onlooker) or a passenger on the train (participant in the scenario). Phrases in italics indicate variations between conditions, *onlooker/participant*:

> Peter is *taking his daily walk near the train tracks when he sees a runaway train approaching with no driver/a passenger on a train whose driver has just shouted that the brakes have failed, and who then fainted of shock.* The train is moving so fast that anyone it hits will die immediately. There are five people working on the main track. It is obvious that they will not be able to get off the track in time and, if nothing is done, they will be killed.
>
> The track has a side-track leading off to the left. *Peter is standing next to a lever. If he pulls the lever, that will/Peter can* turn the train onto the side track and the five people on the main track will not die. But a person is working on the side track. If the train goes onto the side track, then the person on the side track will die. Peter is aware of all these facts.
>
> Thus, Peter can pull the lever, in which case the one person will die but the five people will not; or Peter can refrain from pulling the lever, in which case the five people will die but the one person will not.

In the *game show* scenarios, we moved the action to a game show and varied whether the agent was an audience member (onlooker) or a contestant (a participant). Phrases in italics indicate variations between conditions (*onlooker/participant*) and *large/small* loss:

> Peter is a *member of the studio audience watching/contestant on* a game show. Five contestants have each earned *£200,000/ £10* prize money by answering questions over several rounds, and their tokens are nearing the winning side of the game board. A ball is suddenly released and is rolling toward the tokens of the five contestants and, if nothing is done, they will be knocked out of the game and lose their prize money.
>
> *Peter sees that a button on his armrest has just lit up to indicate that he has been randomly selected by computer to take part in the show.* Peter has the option to press *the/a* button and knock the ball onto another path. But another contestant, who has also earned *£200,000/ £10* prize money, has a token on the new path and will be knocked out of the game and lose his prize money. *Whether or not he presses the button will not affect Peter's winnings.* Peter is aware of all these facts.

Procedure

Subjects were randomly allocated to read only one of the six scenarios. After reading the scenario subjects were asked:

(1) Is it morally wrong for Peter to *turn the train/press the button?* (Yes/No) and to rate the moral right or wrongness of the action on a seven point scale (-3 *Definitely wrong* to +3 *Definitely right*).
(2) Should Peter *pull the lever/press the button?* (Yes/No).
(3) To what extent is it Peter's responsibility to *turn the train/press the button*?, rated on a seven point scale (-3 *Not at all* to +3 *Totally*).
(4) Assuming that Peter *pulled the lever/pressed the button*, to what extent do you agree with the following statements:

- Peter intended that the *person on the side track would die/contestant with the token on the new path would lose their prize money*
- Peter is to blame for the death of the *person on the side-track/loss of the prize money of the contestant with the token on the new path*
- Peter caused the *death of the person on the side-track/loss of the prize money of the contestant with the token on the new path*
- Peter intentionally *killed the person on the side-track/lost the prize money of the contestant with the token on the new path*

These were all rated on a rated on a seven point scale (1 *strongly disagree*, to 7 *strongly agree*).

(5) How believable is this scenario? Rated on a seven point scale (1 *Not at all believable*, to 7 *Completely believable*).

RESULTS

Some subjects did not answer all the survey questions. We did not want to create a sample selection bias by only analyzing data from subjects who completed the whole experiment, so the degrees of freedom in the analyses vary depending on how many subjects responded to the question being analyzed.

Believability of contexts

Our aim of using the game shows to provide a more realistic context was successful. A Two-Way ANOVA revealed a significant main effect of context on judgments of how believable the scenario was: $F_{(2,\ 1134)} = 51.96$, $p < 0.001$, $\eta_p{}^2 = 0.084$. Tukey *post-hoc* tests revealed that the train context ($M = 3.16$) was significantly less believable than the two game show contexts (large game show $M = 4.18$, small game show $M = 4.45$), both $p < 0.001$. On average, subjects rated all the game show scenarios as believable and the train scenarios as unbelievable. There was also a significant main effect of role, with the onlooker scenarios rated as less believable ($M = 3.67$) than the participant scenarios ($M = 4.16$): $F_{(1,\ 1134)} = 19.89$, $p < 0.001$, $\eta_p{}^2 = 0.017$. There was no significant interaction.

Ratings of "how believable is this scenario?" had a negligible correlation with moral judgment, $r_{(1137)} = 0.062$, $p = 0.037$.

Moral judgments

A Two-Way ANOVA showed a significant interaction effect of context and role on rightness judgments: $F_{(2,\ 1175)} = 7.98$, $p < 0.001$, $\eta_p{}^2 = 0.013$ (See **Figure 1** for the mean ratings in each scenario). There was no main effect of context, $F_{(2,\ 1181)} = 2.33$, $p = 0.098$, or of role, $F_{(1,\ 1181)} = 2.76$, $p = 0.097$.

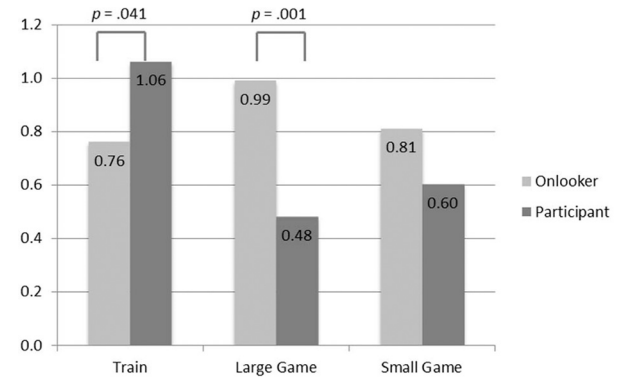


FIGURE 1 | Mean moral rightness ratings (−3 Definitely wrong to +3 Definitely right).

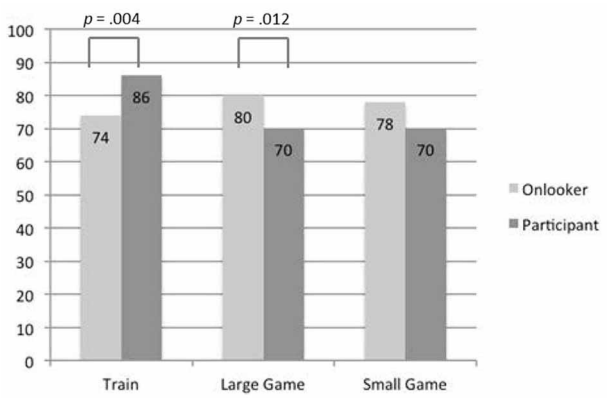


FIGURE 2 | Percentage of subjects who judged that Peter should pull the lever/press the button.

A simple effects analysis across the contexts shows that there was a difference in the way that subjects rated the action of the participant, $F_{(2,\ 1175)} = 9.44$, $p < 0.001$, but not of the onlooker, $F_{(2,\ 1175)} = 1.34$, $p = 0.261$. In the train context, subjects rated the action as more right if the actor was a participant than an onlooker, $F_{(1,\ 1175)} = 4.18$, $p = 0.041$, but in the large game show context this effect was reversed and the actions of a participant (contestant) were rated as less right than those of an onlooker (audience member), $F_{(1,\ 1175)} = 12.78$, $p < 0.001$. There was no effect of role in the small game show context $F_{(1,\ 1175)} = 1.93$, $p = 0.165$.

Judgments of whether or not Peter *should* pull the lever/press the button differed across the six conditions: $\chi^2_{(5,\ 1184)} = 23.21$, $p < 0.001$, $\varphi_c = 0.14$. These results are summarized in **Figure 2**. Pairwise comparisons show that the difference between onlookers and participants is highly significant in the *train* scenario, $\chi^2_{(1,\ 400)} = 8.51$, $p = 0.004$, $\varphi_c = 0.146$, and the *large game show* scenario, $\chi^2_{(1,\ 419)} = 6.33$, $p = 0.012$, $\varphi_c = 0.123$, but narrowly failed to reach conventional levels of significance in the *small game show* scenario: $\chi^2_{(1,\ 365)} = 3.38$, $p = 0.066$, $\varphi_c = 0.096$. In the train scenarios, more people judged that the participant should take the action than the onlooker, but in the game show scenarios more people thought that the onlooker (audience member) should take the action than the participant (contestant). This is the same pattern as the moral judgments.

Relation of responsibility, causation, intention, intentionality, and blame to moral judgment

If we look at how each of the factors varies with role and context, using Two-Way ANOVAs, then we find that subjects gave higher ratings for *caused*, *intentionally*, *intended*, and *blame* in the game show scenarios than in the train, all $p < 0.001$ (see **Table 1**). [The same pattern of results is obtained from a regression analysis. We present partial correlation coefficients in order to make it clear that we make no claims about the direction of causality, which is contested. For opposing views about the direction of causality see Hauser et al. (2007) and Knobe (2010).] There are no significant differences for *responsible*, and no effect of role on any of these factors, or any interaction effects.

Table 1 | Mean (and standard deviation) of factor ratings in the scenarios.

Context	Factor				
	Intended	**Intentionally**	**Caused**	**Blame**	**Responsible**
Train	3.10 (2.17)	2.69 (2.07)	4.31 (2.09)	3.25 (2.16)	0.06 (1.93)
Large game show	3.68 (1.97)	3.79 (1.95)	4.97 (1.86)	4.10 (1.99)	0.07 (1.95)
Small game show	3.74 (1.95)	3.85 (1.98)	5.20 (1.80)	4.37 (1.98)	−0.14 (1.97)

Notes: Responsible rated on a seven point scale (−3 Not at all to +3 Totally), others all rated on a rated on a seven point scale (1 strongly disagree, to 7 strongly agree). For all factors apart from responsibility, means for the game show contexts are different from the mean for the train, p < 0.001. The means for each factor do not differ between the large and small game show contexts.

Table 2 | Partial correlations of the five factors with moral rightness rating.

Intended	**Intentionally**	**Caused**	**Blame**	**Responsible**
−0.038	0.009	0.056	−0.164	0.373
$p = 0.206$	$p = 0.769$	$p = 0.062$	$p < 0.001$	$p < 0.001$

When we look at the partial correlation coefficients, controlling for the presence of the other variables, we find that only *blame* and *responsible* are correlated with the moral judgment of rightness (see **Table 2**), but *intended, intentionally* and *caused* are all correlated with *blame* (see **Table 3**).

DISCUSSION

We found a difference in moral judgment associated with the role of the actor in the scenario, who was the target of the judgment, but the direction of this difference changed depending on the context. In the standard train context, subjects judged that it was more morally right for a passenger, who was already involved in the situation, to turn the train than a bystander, who was an onlooker just passing by. In the game show contexts, it was judged more right for audience members, who were onlookers, than players, who were participating in the quiz, to press the button.

Table 3 | Partial correlations of the five factors with each other.

Factors	Intended	Intentionally	Cause	Blame	Responsible
Intentionally	0.469, $p < 0.001$	1.00			
Caused	−0.064, $p = 0.031$	0.272, $p < 0.001$	1.00		
Blame	0.144, $p < 0.001$	0.121, $p < 0.001$	0.467, $p < 0.001$	1.00	
Responsible	0.038, $p = 0.197$	−0.010, $p = 0.727$	−0.053, $p = 0.073$	0.055, $p = 0.066$	1.00

Subjects' judgments of what the person in the scenario ought to do followed the same pattern as their moral judgments.

Subjects ascribed a higher degree of causation, intentionality, intention, and blame for the harm in the game show than in the train context. When we tested for relationships between each of these factors and moral judgment, whilst controlling for the other factors, we found that blame was the only factor that both correlated with moral judgment and differentiated the game show from the train context. In turn, the increased blame was related to the actors in the game show being rated higher than those in the train scenario on whether they caused the harm, intended the harm, and brought about the harm to the one intentionally. Hence our data suggest that the relation between moral intuitions and intentionality found by Sinnott-Armstrong et al. (2008) and between moral intuitions and intention, proposed by Mikhail (2007) and found by Hauser et al. (2007), is mediated by differential placing of blame.

Responsibility for taking the action correlated with moral judgments when the other factors were controlled. However, responsibility ratings did not differ between the train and the game show contexts. Thus, Thomson's (1985) suggestion that moral intuitions are related to placing of responsibility is supported, although it does not seem that participants have a greater responsibility to take the action than onlookers.

Causes of the reversal

There are two salient differences between the train and the game show scenarios: we changed the context from a train to a game show, and the harmful consequence from death to an economic loss. We think that the reversal of the role effect relates to the change in context, rather than the use of economic harms.

Other studies have replicated trolley results using economic harms. Standard patterns of judgments are seen when economic harms are substituted for mortal harms in hypothetical Sidetrack and Footbridge scenarios (Gold et al., 2013), and when those judgments are being made about decisions in real Sidetrack and Footbridge scenarios, involving small economic harms (Gold et al., submitted). Therefore the reversal of the role effect in our hypothetical scenarios seems likely to be related to the change in context, rather than the substituting of economic harm for mortal harms.

Nor do we think that the reversal we found is related to the fact that the game show winnings have been acquired during a show that has not yet ended. One obvious thought is that game show winnings are "funny money," regarded as not really in the possession of the winner, at least for the duration of the show. However Post et al. (2008) analyzed the behavior of contestants on the television show "Deal or No Deal?" and found that it was consistent with a prospect theory model where decision-makers incorporate expected winnings into their reference point (although the adjustment of the reference point was lagged). This result was not limited to the high stake television game show. It was replicated in classroom experiments with stakes that were 1000 and 10,000 times lower than those on TV. What contestants regard as their current wealth is based on their expectations of how much they will take home, and diminished expectations of winnings represent losses.

The change in context may have affected the causal model that subjects used when representing the problems to themselves—it certainly affected their ascriptions of causation and intentionality—and changing the causal model may affect moral judgments (Spranca et al., 1991; Pizarro et al., 2003; Waldmann and Dieterich, 2007). Whether causation really varies between the train and the game show contexts is a matter for debate. The scenarios were designed so that the explicit causal structures are the same in both contexts. However, the two contexts may have evoked different background assumptions, for example that, in a game show, there are humans involved in running the show who have a causal role in the outcomes and who may bear some blame, whereas in the train context there is no obvious person who is causally responsible or to blame for the malfunction of the train.

The two contexts may also have differed with respect to which agents are perceived to have the right to cause the harm: the participant (passenger) in the train context but the onlooker (audience member) in the game show. People who say that they would not turn the trolley give reasons including not having the right to decide and not wanting to be responsible for someone's death (Gold et al., 2014). Similarly, people who say that they would not vaccinate their child if there was a risk of death cite being responsible for any negative consequence of the action (Ritov and Baron, 1990). (Being responsible for the bad consequence is subtly different from the question we asked, about being responsible for acting, and having the right to act is clearly different from having the responsibility—or duty—to act). Having rights and responsibilities can be connected to the social roles we occupy (Baron, 1996), so the right thing to do in dilemmas with similar structures can be sensitive to context.

STUDY 2

In our second study we operationalized the small game show in a laboratory setting. We conducted a quiz, which subjects either took part in (players) or watched (audience), with monetary prizes for all players who correctly completed more than 15 out of 20 questions. Once at least six players had answered enough questions correctly to collect prizes, we paused the quiz and threatened to knock out five of them, who would lose their winnings. The actor had to decide whether to press a button to keep them in, with the side-effect that we would knock one, different player

out of the quiz, who would lose his or her winnings. We varied whether the actor was a player or an audience member.

This enabled us to investigate whether the role of the actor, who was the target of judgment, would affect judgments in a real life scenario (Target Role: *target player* vs. *target audience*). We also varied the role of the subjects who were making the judgments, (Subject Role: *player* vs. *audience*).

Since we had to have some subjects making the decisions that were being judged, we were also able to observe behavior and to compare the judgments of *actors*, who made decisions and judged the morality of their own decision, with those of *observers*, who made judgments about the action of an actor (Decision Making Power: *actor* vs. *observer*). Actors always made judgments about their own action, so we did not cross subject role and target role for actors.

In Study 1, our questions were all about a third person ("should Peter/a passenger press the button?"). In Study 2, the actors were asked about their own actions and the observers were asked about a third person ("should the player/audience member press the button?"). Hence the nearest equivalent to the difference investigated in Study 1 is when the subject is an observer and the target role is varied, target player vs. target audience.

METHODS

Subjects

There were 202 subjects, 105 men, and 97 women. They were aged between 18 and 56 years ($M = 22.02$, $SD = 6.11$). Subjects were recruited through the University of Leicester's online e-bulletin, which goes out to staff and students. They were tested in groups of 35–40.

Procedure and materials

Subjects sat at computer terminals in one large room and took part in a quiz show (see **Figure 3**). We randomly selected 60% of the subjects to be players, taking part in a general knowledge quiz, and they were assigned pseudonyms. The other 40% were the audience, watching the quiz on their screens. The audience saw the questions in real time and watched the progress of avatars, representing the players, moving across the screen. Players who answered fifteen questions correctly entered the winning zone. Subjects were told that any player who was in the winning zone at the end of the quiz would get £10, and any player who correctly answered nineteen or twenty questions would receive £15. At the end of the experiment, players were paid their winnings, or a £5 show-up fee if they won nothing. Audience members were paid £5 for their participation.

Once six players had entered the winning zone, the quiz stopped. The six players in the zone received a screen message saying "please wait." These players took no decisions and thus these 36 subjects provided no further data to the experiment. Other players received a screen message, whose content depended on the condition that they were in.

Actors (both players and audience members) received the following message:

> Five of the players who are in the winning zone are about to be knocked out of the game by the experimenter and will each lose their £10 cash winnings. You can stop the five from losing their winnings by pressing the button below. However, in that case the experimenter will knock out a different player who is in the winning zone, and the one player will lose his/her £10 cash.

Those actors who were players were also told:

> Whether or not you press the button won't affect your winnings. If you are in the winning zone, then you are not one of the players who is affected by this decision.

Observers received the following message, phrases in italics varied, depending on whether the actor whose behavior was being judged was a player or an audience member:

> Five of the players who are in the winning zone are about to be knocked out of the game by the experimenter and will each lose their £10 cash winnings. *Another of the players/An audience member* is being given the option of pressing a button to stop the five from losing their winnings. However, in that case the experimenter will knock out a different player who is in the winning zone, and the one player will lose his/her £10 cash.

In addition, those who were judging a player were also told:

> Whether or not the player presses the button won't affect his/her winnings. If s/he is in the winning zone, then s/he is not one of the players who is affected by the decision and s/he knows this.

Actors then had 60 s to decide whether or not to push the button. Observers were asked how strongly they agreed with the statement: The *player/audience member* should press the button, rated on a nine point scale (1 *Strongly disagree* to 9 *Strongly agree*).

Subjects were then asked to indicate how wrong or how right it would have been to press the button, on a scale from 1 (*Definitely wrong*) to 9 (*Definitely right*).

At the beginning of the experiment, subjects had been told that "In this experiment some decisions will affect other subjects' payments and some will not," and one randomly selected actor's decision was implemented to see who got knocked out, the one player or the five.

RESULTS

Among actors, the decision to press the button or not was unaffected by whether the person given the choice was a player (78.57% pressed it) or audience member (76.67% pressed it), $\chi^2_{(1,\ 58)} = 0.030$, $p = 0.862$, $\varphi_c = 0.023$. Thus it seems that being a part of the quiz did not increase the proportion of people willing to press the button compared to the people who were merely watching it.

Observers' judgments of whether the actor should press the button were affected by their own roles, as player or audience member (see **Figure 4**). We examined the mean ratings of whether the observers thought that the actor should press the button (1 *strongly disagree* to 9 *strongly agree*) as the dependent variable in a Two-Way ANOVA with Subject Role (*player* vs. *audience*) and Target Role (*target player* vs. *target audience*) as independent variables. There was a main effect of Subject Role, with audience members agreeing more strongly that the

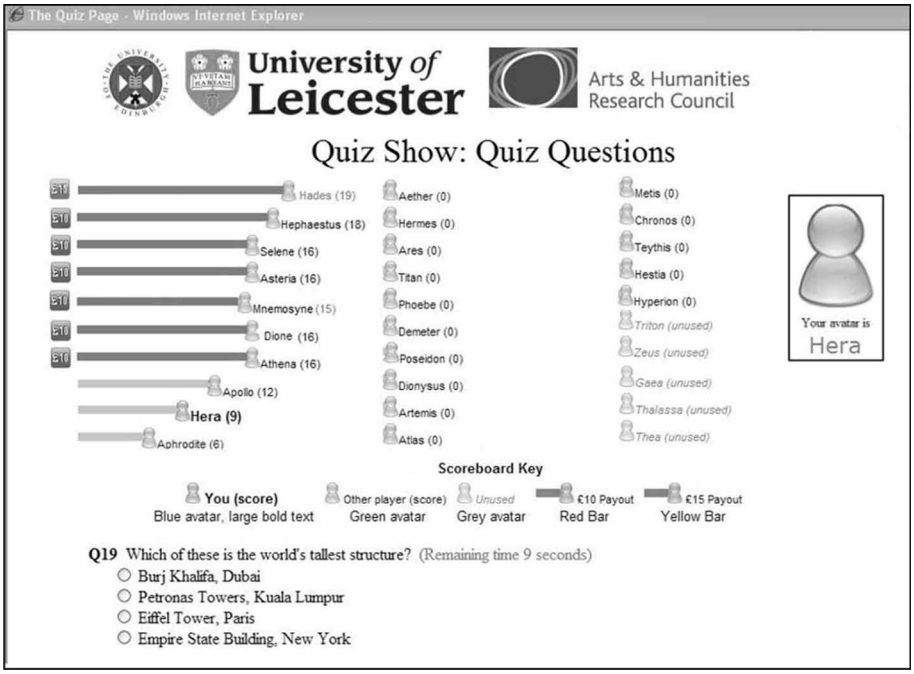


FIGURE 3 | Example of a subject's monitor displaying the questions and the progress of the quiz show subjects.

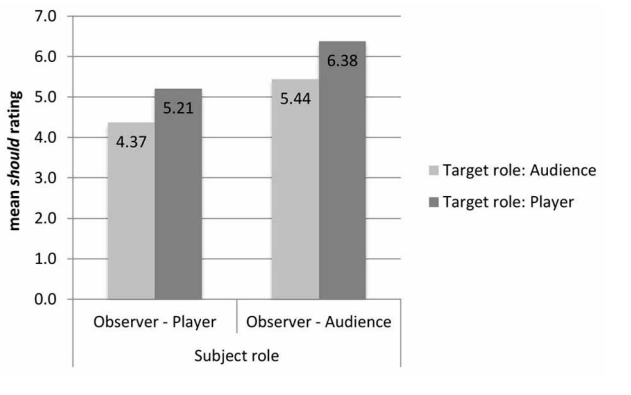


FIGURE 4 | Mean rating of "The player/audience member should press the button" (1 Strongly disagree to 9 Strongly agree).

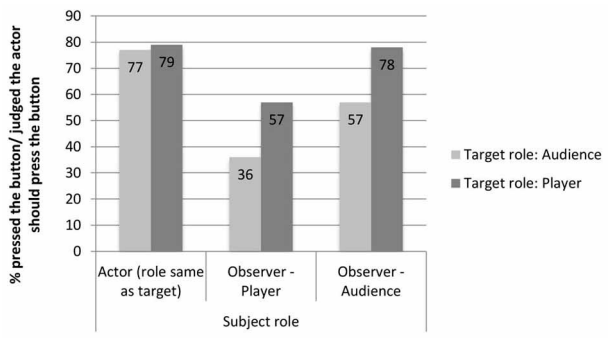


FIGURE 5 | Percentage of actors who pressed the button, and percentage of observers who judged that their target should press the button, broken down by the role of the subject and the role of the target.

actor should press the button ($M = 5.92$) than the players ($M = 4.79$), $F_{(1,\ 104)} = 5.12$, $p = 0.026$, $\eta_p^2 = 0.047$. On average, audience members believed that the actor should press the button (mean rating above 5), but players did not (mean rating below 5). Regarding the Target Role, there was a trend for the player to be judged higher than the audience member (5.80 vs. 4.91), $p = 0.075$, $\eta_p^2 = 0.030$. Our subjects believed, on average, that the player should press the button (mean rating above 5), but that the audience member should not (mean rating below 5). Note that this trend is in the opposite direction of the effect we found in Study 1. There was no interaction between the two factors.

If we group the observers into those who judged that the actor should not take the action (those who gave a rating from 1 to 4) and those who judged that s/he should (rating from 6 to 9), we can more easily compare the data to both the actions of the actors (see **Figure 5**).

Actors were more likely to take action than observers were to judge that they should. The only condition where judgments about what should be done corresponded to what was actually done was that of the audience members whose judgment targeted a player.

There was also an effect due to the role of the person making the judgment. Observers' judgments of whether or not the target should press the button differed across the four conditions: $\chi^2_{(3,\ 91)} = 8.07$, $p = 0.044$, $\varphi_c = 0.298$. The observer's

judgments differ depending on the role of the subject, (67.4% of the observers who are audience members say the target should press the button and 46.7% of the observer players say the target should press the button), $\chi^2_{(1,\ 91)} = 3.99$, $p = 0.046$ $\varphi_c = 0.209$, and depending on the target of judgment, (46.7% of the observers think that the audience members should press the button and 67.4% think that players should press the button), $\chi^2_{(1,\ 91)} = 3.99$, $p = 0.046$, $\varphi_c = 0.209$. This grouping corroborates the pattern found in the ratings: there was a clear reversal of the target-role effect found in the hypothetical game shows in Study 1. In Study 1, subjects were more likely to say that the audience members should press the button, whereas in Study 2, the observers (whose positions correspond most closely to the subjects in Study 1) are more likely to say that the player should take action than the audience member.

The rating of how right or wrong pressing the button was did not vary according to whether the subject was an actor or an observer ($M = 4.22$ vs. 4.45), $F_{(1,\ 160)} = 0.13$, $p = 0.722$, or with Subject Role, audience or player, ($M = 4.10$ vs. 4.65), $F_{(1,\ 160)} = 2.71$, $p = 0.101$, or Target Role, audience or player, ($M = 4.48$ vs. 4.27), $F_{(1,\ 160)} = 1.19$, $p = 0.276$, nor was there any interaction between these factors. In every condition, the average right-wrong judgment fell on the "wrong" side of the scale yet, in three out of four of the observer conditions, a majority of subjects judged that the actor should press the button, and a large majority of actors pressed the button.

DISCUSSION

In Study 2, we operationalized the small game show from Study 1, and the target-role effect in "should" judgments reversed back to being in the same direction as the judgments in the hypothetical train context: when the target was a player more subjects thought that s/he should press the button than when the target was an audience member. Our results are consistent with other evidence that real moral decisions can dramatically contradict moral choices made in hypothetical scenarios (FeldmanHall et al., 2012b).

A key difference between our real and realistic scenarios is that actually being in the scenario may have evoked a "hot" affective state whereas contemplating the same hypothetical scenario is done in an affectively "cold" state. Differences in affective states between real and hypothetical scenarios could cause judgments and behavior to be different (Kühberger et al., 2002; Kang and Camerer, 2013). People are probably not even aware that their judgments would differ in real and hypothetical scenarios because there is a "hot/ cold empathy gap," where people mispredict the effect of their affective state on their preferences and behavior (Loewenstein, 2005). Yet, in a real task, manipulating whether participants are in "hot" or "cold" states affects behavior, with the "hot" version being associated with more risk taking and poorer information use (Figner et al., 2009).

The importance of affective states is supported by neuroscientific evidence. Real and hypothetical moral decisions differentially recruit neural circuitry, with hypothetical moral decisions eliciting activity in neural circuits that are involved in imagination, whilst real moral decisions activate the amygdala, which is crucial for social and affective processes (FeldmanHall et al., 2012a). There is also increased activity in the amygdala when subjects are presented with stories that narrate their own intentional violation of social norms, compared to violations by others; this has been linked to enhanced emotional responses (Berthoz et al., 2006).

Others have stressed the importance of emotional reactions in trolley problems (e.g., Greene et al., 2001) and "hot" affect may connect our outlandish and real scenarios. The outlandish trolley scenario may elicit a strong emotional response because the hypothetical outcomes involve deaths; the real scenario may evoke an emotional response because the small harms will actually occur. So both the outlandish and the real scenarios may have provoked a more emotional response than the realistic scenarios. Thus we observed similar patterns of responses in the outlandish and the real scenarios and a different pattern in the realistic scenarios.

There was also a difference in judgments depending on the role of the subject making the judgment: audience members were more likely to judge that the actor should press the button. Audience members and players might have differentially empathized with the one player who risked being knocked out, with players being more likely to think "what if it were me?" Interestingly, when observers judge people in the same role as themselves—when players judge players and audience members judge audience members—57% of both groups think that the actor should press the button. It is when these two groups judge people from a different role that stark differences appear. When audience members judge players 78% of them think the player should press the button, a figure that matches almost precisely the number of actors who actually do take action. In contrast, only 36% of the players think that an audience member should press the button to save the five from losing their money, thus indicating that the majority of players feel that the audience members should stay out of the situation and not intervene.

Actors consistently pressed the button, and more actors pressed the button than observers said should press the button. We did not ask actors for their judgment about what they should do, as it risked merely eliciting self-justificatory answers. If the observers' judgments are indicative of what actors thought they should do, then many actors pressed the button despite thinking that they ought not to. This is a case of weakness of will. Alternatively, if actors acted in line with their judgments about what they ought to do, then having the power to make a decision affects one's judgment about what ought to be done. In either case, it appears that asking an observer what should be done gets different results from observing actual actions.

Despite the difference in opinions about what should be done amongst observers, there are no differences in moral judgments between the groups in Study 2. Different patterns of hypothetical choice and moral judgments have also been found by Tassy et al. (2013), who hypothesize that this occurs because choice and judgment are the results of different psychological processes; and different patterns of actual choice and moral judgments have been found by Gold et al. (submitted), who suggest that their subjects found that the normatively relevant factors for whether or not to press the button were not exhausted by its moral right and wrongness. There may be pragmatic factors in play.

GENERAL DISCUSSION

We found differences in moral judgments between outlandish and realistic hypothetical scenarios, and between judgments made in hypothetical scenarios vs. the same scenarios operationalized in real-life. Of course, showing that there are differing responses cannot tell us which responses are "correct" or which type of scenarios we should study (Elqayam and Evans, 2011). But we can outline some of the advantages and disadvantages of each approach.

Researchers may choose to use outlandish artificial dilemmas, rather than realistic ones, in order to isolate the dimensions that are of theoretical interest (Hauser et al., 2007). Real life scenarios are usually complex, so isolating dimensions of interest generally necessitates using outlandish scenarios. Some researchers see subjects' lack of familiarity with the outlandish scenarios as a further point in their favor, because it removes some of the social and personal factors that might otherwise influence responses (Hauser et al., 2007). But both of these supposed benefits are contested, particularly when dilemmas are used in ethics. There is a move, especially in medical ethics, to see moral dilemmas as occurring within a broader narrative, so their resolution requires moral imagination and a more holistic engagement with all the features of the case (Hunter, 1996; London, 2001). There are also arguments that we can be most sure of our moral judgments when we contemplate complicated and familiar cases: either particular paradigm cases, such as landmark legal cases (Jonsen and Toulmin, 1988), or familiar situations (Woodward and Allman, 2007).

Real and hypothetical dilemmas may put subjects in different affective states (Kühberger et al., 2002; Kang and Camerer, 2013). There is disagreement whether subjects should be in "hot" or "cold" states when moral judgments are elicited. Real-life moral cognition is hot cognition and, if hot and cold judgments differ, especially if they involve different brain systems, it follows that psychological studies of moral cognition would benefit from being done in ecologically valid settings (Casebeer, 2003; Moll et al., 2005). However, when judgments are used for philosophical purposes, it has been argued that we should be wary of judgments that are driven by " 'alarm bell' emotion" (Greene, 2007, p.63), which suggests privileging "cold" judgments.

Researchers should bear in mind that whether scenarios are outlandish, realistic, or real may affect moral judgments. But which type of scenarios is most appropriate to use may depend on the nature and purpose of the study. Furthermore, a complete understanding of the significant differences reported in our experiments will, of course, require a great deal more research, and the potential explanations are myriad. It is even possible, following a suggestion made by Skinner (1985) in a generalized critique of cognitive science, that the differences could be explained by people's application of patterns of behavior learnt under contingencies of reinforcement in analogous situations experienced in everyday life. However, such purely behavioral explanations are bound to exist alongside interpretations in cognitive and ethical terms.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from the Arts and Humanities Research Council grant AH/H001158/1, from the University of Leicester for granting study leave to Briony Pulford, and from the European Research Council who supported Natalie Gold during the revisions to this article under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 283849. We thank Jamie Lien for helpful input, Kevin McCracken for programming software for Study 2, and our research assistants Manisha Chauhan and Catherine Lawrence.

REFERENCES

Andreoni, J., Brown, P. M., and Vesterlund, L. (2002). What makes an allocation fair? Some experimental evidence. *Games Econ. Behav.* 40, 1–24. doi: 10.1006/game.2001.0904

Andreoni, J., and Miller, J. (2002). Giving according to GARP: an experimental test of the consistency of preferences for altruism. *Econometrica* 70, 737–753. doi: 10.1111/1468-0262.00302

Baron, J. (1996). "Do no harm," in *Codes of Conduct: Behavioral Research into Business Ethics*, eds D. M. Messick and A. E. Tenbrunsel (New York, NY: Russell Sage Foundation), 197–213.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games Econ. Behav.* 10, 122–142. doi: 10.1006/game.1995.1027

Berthoz, S., Grezes, J., Armony, J., Passingham, R., and Dolan, R. (2006). Affective response to one's own moral violations. *Neuroimage* 31, 945–950. doi: 10.1016/j.neuroimage.2005.12.039

Bolton, G., and Ockenfels, A. (2000). ERC: a theory of equity, reciprocity, and competition. *Am. Econ. Rev.* 90, 166–193. doi: 10.1257/aer.90.1.166

Casebeer, W. D. (2003). Moral cognition and its neural constituents. *Nat. Rev. Neurosci.* 4, 840–847. doi: 10.1038/nrn1223

CNN U.S. (2003). Runaway freight train derails near Los Angeles. *CNN U.S.* Available online at: http://articles.cnn.com/2003-06-20/us/train.derails_1_derailment-freight-cars-runaway-freight-train?_s=PM:US

Cushman, F. A., Young, L., and Hauser, M. D. (2006). The role of conscious reasoning and intuition in moral judgments: testing three principles of harm. *Psychol. Sci.* 17, 1082–1089. doi: 10.1111/j.1467-9280.2006.01834.x

Elqayam, S., and Evans, J. S. B. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868. doi: 10.1162/003355399556151

Fehr, E., and Schmidt, K. M. (2006). "The economics of fairness, reciprocity and altruism – experimental evidence and new theories," in *Handbook of the Economics of Giving, Altruism and Reciprocity,* Vol. 1, eds S.-C. Kolm and J. M. Ythier (North Holland: Elsevier), 615–691. doi: 10.1016/S1574-0714(06)01008-6

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., and Mobbs, D. (2012a). Differential neural circuitry and self-interest in real vs. hypothetical moral decisions. *Soc. Cogn. Affect. Neurosci.* 7, 743–751. doi: 10.1093/scan/nss069

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., and Dalgleish, T. (2012b). What we say and what we do: the relationship between real and hypothetical moral choices. *Cognition* 123, 434–441. doi: 10.1016/j.cognition.2012.02.001

Figner, B., Mackinlay, R. J., Wilkening, F., and Weber, E. U. (2009). Affective and deliberative processes in risky choice: age differences in risk taking in the Columbia Card Task. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 709–730. doi: 10.1037/a0014983

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Rev.* 5, 5–15.

Gigerenzer, G., Todd, P., and the ABC Research Group. (1999). *Simple Heuristics that Make us Smart*. New York, NY: Oxford University Press.

Gold, N., Colman, A. M., and Pulford, B. D. (2014). Cultural differences in response to real-life and hypothetical trolley problems. *Judgm. Decis. Mak.* 9, 65–76. Available online at: http://journal.sjdm.org/12/121101/jdm121101.pdf

Gold, N., Pulford, B. D., and Colman, A. M. (2013). Your money or your life: comparing judgments in trolley problems involving economic and emotional harms, injury and death. *Econ. Philos.* 29, 213–233. doi: 10.1017/S0266267113000205

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108. doi: 10.1126/science.1062872

Greene, J. D. (2007). "The secret joke of Kant's soul," in *Moral Psychology, The Neuroscience of Morality: Emotion, Disease, and Development*, Vol. 3, ed W. Sinnott-Armstrong (Cambridge, MA: MIT Press).

Hauser, M. D., Cushman, F. A., Young, L., Kang-Xing Jin, R., and Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind Lang.* 22, 1–21. doi: 10.1111/j.1468-0017.2006.00297.x

Hunter, K. M. (1996). Narrative, literature, and the clinical exercise of practical reason. *J. Med. Philos.* 21, 303–320. doi: 10.1093/jmp/21.3.303

Jonsen, A. R., and Toulmin, S. (1988). *The Abuse of Casuistry: A History of Moral Reasoning.* Berkeley, CA: University of California Press.

Kamm, F. M. (1989). Harming some to save others. *Philos. Stud.* 57, 227–260. doi: 10.1007/BF00372696

Kamm, F. M. (1996). *Morality, Mortality, Volume II: Death and Whom to Save from it*. New York, NY: Oxford University Press.

Kang, M. J., and Camerer, C. F. (2013). fMRI evidence of a hot-cold empathy gap in hypothetical and real aversive choices. *Front. Neurosci.* 7:104. doi: 10.3389/fnins.2013.00104

Knobe, J. (2010). Action trees and moral judgment. *Top. Cogn. Sci.* 3, 555–578. doi: 10.1111/j.1756-8765.2010.01093.x

Knutson, K. M., Krueger, F., Koenigs, M., Hawley, A., Escobedo, J. R., Vasudeva, V., et al. (2010). Behavioral norms for condensed moral vignettes. *Soc. Cogn. Affect. Neurosci.* 5, 378–384. doi: 10.1093/scan/nsq005

Kühberger, A., Schulte-Mecklenbeck, M., and Perner, J. (2002). Framing decisions: hypothetical and real. *Organ. Behav. Hum. Decis. Process.* 89, 1162–1175. doi: 10.1016/S0749-5978(02)00021-3

Loewenstein, G. (2005). Hot-cold empathy gaps and medical decision making. *Health Psychol.* 24, S49–S56. doi: 10.1037/0278-6133.24.4.S49

London, A. J. (2001). The independence of practical ethics. *Theor. Med. Bioeth.* 22, 87–105. doi: 10.1023/A:1011403909450

Mikhail, J. (2007). Universal moral grammar: theory, evidence, and the future. *Trends Cogn. Sci.* 11, 143–152. doi: 10.1016/j.tics.2006.12.007

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., and Grafman, J. (2005). Opinion: the neural basis of human moral cognition. *Nat. Rev. Neurosci.* 6, 799–809. doi: 10.1038/nrn1768

Nadelhoffer, T., and Feltz, A. (2008). The actor-observer bias and moral intuitions: adding fuel to Sinnott-Armstrong's fire. *Neuroethics* 1, 133–144. doi: 10.1007/s12152-008-9015-7

Pizarro, D. A., Uhlmann, E., and Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *J. Exp. Soc. Psychol.* 39, 653–660. doi: 10.1016/S0022-1031(03)00041-6

Post, T., Van den Assem, M. J., Baltussen, G., and Thaler, R. H. (2008). Deal or no deal? Decision making under risk in a large-payoff game show. *Am. Econ. Rev.* 98, 38–71. doi: 10.1257/aer.98.1.38

Pulford, B. D., Colman, A. M., and Gold, N. (2012). "Investigating the effects of framing in trolley problems," in *Keynote Paper Presented at the Experiments in Ethical Dilemmas Workshop* (London).

Quinn, W. S. (1989). Actions, intentions, and consequences: the doctrine of doing and allowing. *Philos. Rev.* 98, 287–312. doi: 10.2307/2185021

Ritov, I., and Baron, J. (1990). Reluctance to vaccinate: omission bias and ambiguity. *J. Behav. Decis. Mak.* 3, 263–277. doi: 10.1002/bdm.3960030404

Rozyman, E., and Baron, J. (2002). The preference for indirect harm. *Soc. Justice Res.* 15, 165–184. doi: 10.1023/A:1019923923537

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., and Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an fMRI investigation. *J. Cogn. Neurosci.* 18, 803–817. doi: 10.1162/jocn.2006.18.5.803

Sinnott-Armstrong, W., Mallon, R., McCoy, T., and Hull, J. G. (2008). Intention, temporal order, and moral judgments. *Mind Lang.* 23, 90–106. doi: 10.1111/j.1468-0017.2007.00330.x

Skinner, B. F. (1985). Cognitive science and behaviorism. *Br. J. Psychol.* 76, 291–301. doi: 10.1111/j.2044-8295.1985.tb01953.x

Spranca, M., Minsk, E., and Baron, J. (1991). Omission and commission in judgment and choice. *J. Exp. Soc. Psychol.* 27, 76–105. doi: 10.1016/0022-1031(91)90011-T

Tassy, S., Oullier, O., Mancini, J., and Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Front. Psychol.* 4:250. doi: 10.3389/fpsyg.2013.00250

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *Monist* 59, 204–217. doi: 10.5840/monist197659224

Thomson, J. J. (1985). The trolley problem. *Yale Law J.* 94, 1395–1415. doi: 10.2307/796133

Waldmann, M. R., and Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: intervention myopia in moral intuitions. *Psychol. Sci.* 18, 247–253. doi: 10.1111/j.1467-9280.2007.01884.x

Woodward, J., and Allman, J. (2007). Moral intuition: its neural substrates and normative significance. *J. Physiol. Paris* 101, 179–202. doi: 10.1016/j.jphysparis.2007.12.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 October 2013; paper pending published: 17 December 2013; accepted: 12 January 2013; published online: 30 January 2014.

*Citation: Gold N, Pulford BD and Colman AM (2014) The outlandish, the realistic, and the real: contextual manipulation and agent role effects in trolley problems. Front. Psychol. **5**:35. doi: 10.3389/fpsyg.2014.00035*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on