CAUSAL COGNITION IN HUMANS AND MACHINES

EDITED BY: Andrew Tolmie, Selma Dündar-Coecke and York Hagmayer
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-257-8
DOI 10.3389/978-2-88974-257-8

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: frontiersin.org/about/contact

CAUSAL COGNITION IN HUMANS AND MACHINES

Topic Editors:
Andrew Tolmie, University College London, United Kingdom
Selma Dündar-Coecke, University College London, United Kingdom
York Hagmayer, University of Göttingen, Germany

Citation: Tolmie, A., Dündar-Coecke, S., Hagmayer, Y., eds. (2022). Causal Cognition in Humans and Machines. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88974-257-8

Table of Contents

frontiers in Psychology

PERSPECTIVE
published: 22 January 2020
doi: 10.3389/fpsyg.2020.00003

What Is Causal Cognition?

*Andrea Bender**

Department of Psychosocial Science, SFF Centre for Early Sapiens Behaviour (SapienCE), University of Bergen, Bergen, Norway

OPEN ACCESS

Edited by:
York Hagmayer,
University of Göttingen,
Germany

Reviewed by:
Wolfgang Schoppek,
University of Bayreuth,
Germany
Ion Juvina,
Wright State University,
United States

****Correspondence:***
Andrea Bender
andrea.bender@uib.no

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

Received: *12 August 2019*
Accepted: *03 January 2020*
Published: *22 January 2020*

Citation:
Bender A (2020) What Is
Causal Cognition?
Front. Psychol. 11:3.
doi: 10.3389/fpsyg.2020.00003

While gaining an understanding of cause-effect relations is the key *goal* of causal cognition, its *components* are less clearly delineated. Standard approaches in the field focus on how individuals detect, learn, and reason from statistical regularities, thereby prioritizing cognitive processes over content and context. This article calls for a broadened perspective. To gain a more comprehensive understanding of what is going on when humans engage in causal cognition—including its application to machine cognition—it is argued, we also need to take into account the content that informs the processing, the means and mechanisms of knowledge accumulation and transmission, and the cultural context in which both accumulation and transmission take place.

Keywords: causal cognition, cognitive processes, content, culture, language, knowledge accumulation, knowledge transmission

INTRODUCTION

Causality is the relation between two events, one of which is the consequence (or *effect*) of the other (*cause*). Gaining an understanding of such cause-effect relations is of prime concern for humans, starting in infancy with a drive to explore one's world and test one's assumptions (Gopnik et al., 1999; Muentener and Bonawitz, 2017). Indeed, the ability to attain causal understanding and harness it for diagnoses, predictions, and interventions is so advantageous that it has been considered the main driving force in human evolution (Stuart-Fox, 2015; Lombard and Gärdenfors, 2017).

While understanding is arguably the key *goal* of causal cognition, its *components* are less clearly delineated. So, what exactly is causal cognition? Or rather, how should we conceptualize it from a cognitive science point of view? As will be detailed in the next section, a great deal of approaches in this field focuses on the detection of and reasoning from statistical regularities. Taking this rather narrow focus as the starting point, I will advocate a broader perspective on causal cognition, which also factors in its distinctly human characteristics, specifically the crucial roles of content, knowledge transmission, and culture. Implications for the field—including application to machine cognition—will be discussed prior to the conclusion.

PERSPECTIVES ON CAUSAL COGNITION

The preamble for this research topic outlines causal cognition as the ability "to perceive and reason about [...] cause-effect relations."[1] This outline largely reflects what may be seen as the "standard view" in cognitive and social psychology. In the following, this view will be fleshed out, before addressing the dimensions along which it needs to be extended.

The Standard View

Precise definitions of causal cognition are hard to come by. Scholars tend to presume that the term is self-explanatory and hence only mention in passing what they are actually focusing on. Nevertheless, a reasonably reliable impression can be gleaned from the first five publications that pop up when "causal cognition" is entered into Google Scholar (with jointly 1,280 citations in total, as of 12 August 2019, sorted by relevance).

The three publications which come from cognitive and comparative psychology cast causal cognition as the understanding of causal mechanisms (Zuberbühler, 2000; Penn and Povinelli, 2007) and as representations of the causal relation between action and outcome (Dickinson and Balleine, 2000). That is, concealed by the more generic term "causal cognition," the subject of the respective works is actually confined to just a few aspects, each of which has an entire research tradition devoted to it: perception (Michotte, 1963; Saxe and Carey, 2006), learning (Shanks et al., 1996; Gopnik et al., 2004), and reasoning (Blaisdell et al., 2006; Waldmann, 2017).

Social psychologists add attribution, as their topic of core concern, to this range of cognitive processes (Norenzayan and Nisbett, 2000), that is, explanations of social behavior in terms of dispositional and/or contextual factors (Kelley, 1973; Choi et al., 1999). The cognitive and the social tradition essentially differ in terms of the *explanandum*—a change as the outcome of an event or of one's actions, versus an account of why people behave in a certain way—but they both conceptualize causal cognition as consisting of mental processes.

While some scholars emphasize the domain-general nature of these processes, others consider domain boundaries to be relevant for distinguishing different types of causal cognition (Morris and Peng, 1994). And some even argue for the existence of domain-specific modules devoted to reasoning distinctly about physical, biological, and social/psychological events (Leslie, 1994; Spelke and Kinzler, 2007). Domains in this sense are defined by the distinct properties of their key entities and the causal principles accounting for their behavior. Objects in the physical domain, for instance, move when propelled by external forces in line with mechanistic principles, whereas the inhabitants of the biological domain are able to move of their own accord, in line with vitalistic principles. These different principles motivate a conceptual distinction between the constructs of *cause* (as eliciting a physical effect) versus *reason* (as motivating behavior), and between cognitive processes devoted to physical *causation* (like perception and reasoning) versus those devoted to social *agency* (like attribution and ascription of responsibility).

Only one of the five above-mentioned publications, a multidisciplinary compilation of 20 contributions on causal cognition (Sperber et al., 1995), outlines a broader range of perspectives, regarding both the processes and factors involved and the domains considered.

A More Comprehensive View

Some core components of causal cognition, like learning based on statistical regularities, are firmly rooted in our evolutionary past: They are present in non-human animals, they are observable in human infants, and they enabled our ancestors to move out of their original habitat and spread around the globe (Bender, 2019). Even these shared roots, however, do not render causal cognition a uniform phenomenon. Relevant abilities in infants already transcend those of our closest relatives in several ways. Causal cognition in humans is characterized, *inter alia*, by the integration of content information into theory-like representations, with serious implications for processing. This role of content and the means by which it is incorporated will be outlined in more detail in the following.

The Role of Content for Processing

As noted above, the bulk of research on causal cognition focuses on processing while abstracting from content. As one consequence, methods prioritize artificial tasks in laboratory settings, involving toys and other stimuli designed for the very purpose of bearing no similarity to anything with which participants may be familiar (e.g., Gopnik and Sobel, 2000). Confronted with a meaningless pattern of statistical regularities, the participant's task is to diagnose the underlying causal relations. Oddly enough, the very reason for doing so is that content plays such an overwhelming role in human causal cognition that, to be able to isolate the "pure" processes underlying it, detaching these processes from content appears indispensable.

The most abstract form of content is a structural model of the causal relations involved (e.g., whether they constitute a simple chain or a more complex network), and even rats have been shown to form such deeper causal representations, which lead their learning and reasoning (Blaisdell et al., 2006). When available, knowledge and beliefs on properties of items, on dependencies between them, or even on underlying mechanisms of causation inform these representations of structure. Pieces of knowledge are themselves embedded in mental models of how things work, which in turn guide tool use, decision-making, and problem-solving. For instance, rich knowledge on a domain affords reasoning strategies based on causal mechanisms, rather than category-based induction (Medin and Atran, 2004); and beliefs on causal mechanisms affect not only what, but also how, people decide (Kempton, 1986; Dörner, 1996; Güss and Robinson, 2014). On a higher level still, these various sorts of representations are organized by framework theories. Framework theories are ontological perspectives on the world, enriched with cultural values, that motivate

[1] https://www.frontiersin.org/research-topics/9874/causal-cognition-in-humans-and-machines

interpretations, inferences, and intentions (Bang et al., 2007). They affect, for instance, how information is filed in long-term memory, whether reasoning is biased by typicality and diversity effects, or on which principles domain boundaries are drawn (Medin and Atran, 2004; ojalehto et al., 2017a,b). This need not imply that causal models are uniform or coherent; in fact, apparently incompatible accounts can co-exist in an individual's mind and are selectively accessed depending on contextual cues (Astuti and Harris, 2008; Legare and Gelman, 2008).

In other words, content impacts on processing. If, however, the integration of knowledge and beliefs into theory-like representations is indeed so essential and decisive, accounts of human causal cognition cannot afford to disregard content.

The Role of Knowledge Transmission for Content

A great deal of knowledge about causation can be gleaned from an individual's interactions with the world, and observing statistical regularities may render a reasonably accurate model of causal relations, for instance when trying to diagnose and treat a common cold. Still, accounting for the underlying mechanisms is replete with interpretation and, often enough, pure speculation. The more elaborate such accounts are, the more likely they therefore are to encompass large portions that we simply learned from other people (D'Andrade, 1995).

While learning from others is not an exclusively human ability, the extent to which our species capitalizes on it is indeed unique. Even as young children, humans pay specific attention to social cues (Kushnir et al., 2008), and when copying problem-solving behavior, they "over-imitate," by prioritizing conventional aspects over mechanistic aspects, whether or not the former are causally relevant (Lyons et al., 2007)—a tendency that further increases into adulthood (McGuigan et al., 2011). Humans not only actively seek information, but are also willing to convey it. This willingness arises from our disposition for shared intentionality, for teaching, and for learning from teaching (Tomasello et al., 2005; Csibra and Gergely, 2009).

In contrast to the acquisition of behavioral patterns and action-based problem-solving, teaching is indispensable for the explicit transmission of knowledge, particularly for knowledge on a subject that is as invisible and ephemeral as causality (Waldmann et al., 2006). With language, humans have developed the most powerful tool in the entire animal kingdom for achieving this—a tool that young children already exploit in full when they ask for causal explanations, and persist in requesting more explanations if they are not satisfied with the previous ones (Callanan and Oakes, 1992; Frazier et al., 2009).

Given its key role for knowledge accumulation, the impact of language and its usage on causal cognition should not be underestimated. Sometimes, a linguistic label may be sufficient to serve as a cue for causal assumptions (as is the case with the common cold, which, according to popular belief, is caused by exposure to cold weather). But language use can also affect cognition more subtly, through the ways in which information about causal relations and events is encoded, or in how event descriptions are linguistically prepacked or split into their components (Wolff et al., 2009; Bohnemeyer et al., 2010). For instance, while "the climate is changing" and "humans are changing the climate" both describe the same event, the two linguistic constructions still suggest slightly diverging causal perspectives, one focusing on the event, and the other on the agent. Such modifications of the linguistic framing are able to redirect people's attention to, in this case, event or agent (Fausey et al., 2010); to alter their inferences on causal efficacy (Kuhnmünch and Beller, 2005); to sway their memories of something they themselves observed (Loftus and Palmer, 1974; Fausey et al., 2010); or to affect their assignment of agency, responsibility, and blame (Fausey and Boroditsky, 2010; Bender and Beller, 2017).

In other words, content consists of knowledge that is socially accumulated and transmitted, frequently through explicit teaching using language. If, however, transmission is so crucial for content generation, with the means of transmission affecting causal representations and processing, accounts of human causal cognition cannot afford to disregard the role and the characteristics of the mechanisms involved.

The Role of Culture for Knowledge Transmission

Transmission of knowledge typically takes place within a social context. Social orientations and cultural practices therefore impact on every step of it: the bits and pieces of knowledge transmitted, the means of transmission, and the specific details of the transmission process itself.

As noted above, the bulk of people's knowledge and beliefs is learned from others and hence bears the stamp of the cultural setting in which it emerged and is transmitted. Cultural shaping is amplified insofar as knowledge and beliefs are accumulated over time and integrated into larger models and *framework theories* (Bang et al., 2007). Cultural framework theories not only provide distinct ontological perspectives, and hence endow meaning to the causal accounts of the very same event in notably different ways, but even entail different ways of partitioning the world into domains. The ontological perspective implicit in most Western framework theories, for instance, suggests partitioning into a physical, a biological, and a social-psychological domain, largely based on properties of their key entities and on corresponding principles for agency ascription (Carey, 1996, 2009; Spelke and Kinzler, 2007). The ontological perspective implicit in Amerindian framework theories, by contrast, emphasizes interconnectedness between entities, and hence suggests principles for agency ascription that are grounded in relations rather than properties, and that give rise to domains based on communication and exchange (ojalehto et al., 2017a,b).

As a consequence, causal cognition is infused with culture. People therefore differ in whether they engage in causal considerations on a regular basis (Beer and Bender, 2015), and in how they weigh consequences versus causes (Choi et al., 2003; Maddux and Yuki, 2006). They also differ in the principles in which category and domain boundaries are grounded (ojalehto et al., 2017a,b), and in the concepts that inform their explanations (Beller et al., 2009). Even the biases that affect inferences differ across cultures (Medin and Atran, 2004; Bender and Beller, 2011). Factors contributing to these differences include, among others, the cultural shaping of the settings in which causal cognition occurs; the extent to which socialization patterns and teaching strategies encourage or discourage exploration and requests for explanation; the culture-specific organization of causally relevant knowledge, concepts, and categories; and the language-specific encoding of causal relations in grammatical structure (for reviews, see Bender et al., 2017; Bender and Beller, 2019).

In other words, knowledge transmission is ingrained in culture. If, however, the accumulation and propagation of information is so dependent on cultural practices and institutions, accounts of human causal cognition cannot afford to disregard its cultural fabric.

IMPLICATIONS FOR STUDYING CAUSAL COGNITION IN HUMANS AND MACHINES

While causality might be objective, and our interest in it phylogenetically old, neither of the two is set in stone. As demonstrated by Iliev and colleagues (Iliev and ojalehto, 2015; Iliev and Axelrod, 2016), the extent of our concern with causality has changed over time—even over the course of just one century—and so too has the usage of the corresponding vocabulary and concepts. Here, I argue that our scientific notions of causal cognition can, and in fact must, change as well.

Research on causal cognition has typically focused on how humans gain explanations for what is going on in the world. In so doing, it often reduces causal cognition to a few cognitive processes involved in perception, learning, reasoning, and attribution, which are investigated devoid of content or context. Yet, to achieve a more comprehensive understanding of what is going on when humans engage in causal cognition, we also need to take into account the content that informs the cognitive processing, the means and mechanisms of knowledge accumulation and transmission, and the cultural context in which both accumulation and transmission take place. All of these aspects are unique to, and constitutive of, human causal cognition, and have serious implications for how we study causal cognition in humans and machines.

As a first consequence, we may wish to acknowledge more phenomena as components of causal cognition than just the inferences drawn from patterns of statistical regularities. Included should be, *inter alia*, verbal accounts, principles for categorization, tool use in daily life, problem-solving in complex situations, or judgments of blameworthiness and punishment. Concurrently, the segregation between the physical and the social domain—and hence between causation and agency—should be abolished as arguably culture-specific categorizations.

As a second consequence, we may wish to reconsider the methods we apply for investigating causal cognition. The repertoire of research strategies should be extended beyond philosophical reflections and sterile lab experiments, to also include statistical analyses of linguistic data, in-depth within-culture analyses of cognitive concepts, processes, and changes over time, ethnographic observations, or cross-cultural and cross-linguistic studies (Bender and Beller, 2016). Moreover, stronger efforts should be undertaken to increase the ecological validity afforded by our tools and settings.

A third consequence arises for attempts to model human causal cognition in machines. The recent exceptional progress in the area of artificial intelligence is largely thanks to the harnessing of deep learning for pattern recognition. Basically reflecting the "standard view" of causal cognition, this focus remains on the lowest rung of Pearl's *Ladder of Causation* (Pearl and Mackenzie, 2018) and falls short of resembling human competences. Two of the core ingredients proposed by Lake et al. (2017) for making machines "learn and think like people" include an ability to build causal models and the grounding of learning in intuitive theories of physics and psychology (a kind of developmental "start-up software"). This emphasis on structure and content echoes insights from research on causal cognition in humans and non-human species (Pearl, 2000; Waldmann et al., 2006) and would ensure that most of the shared components of causal cognition are accounted for. Still, for modeling (uniquely) human characteristics, a further step needs to be taken: the implementation of social learning and cultural accumulation of knowledge, possibly enriched by language use (Dennett and Lambert, 2017; Tessler et al., 2017). Learning from others not only requires fewer data and occurs at a higher speed, but is also a key mechanism in diversification. As Clegg and Corriveau (2017) put it: even if the developmental "start-up software" is assumed to be universal, the "software updates" are likely shaped by culture and may over time generate distinct operating systems.

CONCLUSION

To sum up, gaining an understanding of cause-effect relations is an ability in which humans clearly and strikingly outperform any other species. To a great extent, this is due to the fact that in our species, individuals are just not reliant on drawing inferences from observed statistical regularities, each on their own, but are willing and able to share their observations, inferences, and interpretations, to accumulate them over time, and to transmit them to the next generation. The content, which is so crucial in human causal cognition, is a product of culture from the very beginning, rendered possible and profoundly shaped by the fact that humans are a cultural species (Bender and Beller, 2019). While these characteristics of human causal cognition may not be considered relevant when transferring models from humans to machines—or not even desirable in some applications (Livesey et al., 2017)—it would at least be instructive to be aware of them.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was supported by the Research Council of Norway through the SFF *Centre for Early Sapiens Behaviour* (SapienCE), project number 262618. Publication fees are covered by the Library of the University of Bergen.

ACKNOWLEDGMENTS

I would like to thank the fellows of the Research Group "The cultural constitution of causal cognition," funded by the *Center for Interdisciplinary Research* (ZiF) at Bielefeld University, for stimulating discussion, and Sarah Mannion de Hernandez for proof-reading and valuable comments on an earlier draft. This article is based in large parts on collaborative work with my former colleague and partner, the late Sieghard Beller, who got me interested in causal cognition and to whom I remain indebted for inspiration and an exceptionally fruitful exchange of ideas.

REFERENCES

Astuti, R., and Harris, P. L. (2008). Understanding mortality and the life of the ancestors in rural Madagascar. *Cogn. Sci.* 32, 713–740. doi: 10.1080/03640210802066907

Bang, M., Medin, D. L., and Atran, S. (2007). Cultural mosaics and mental models of nature. *Proc. Natl. Acad. Sci. USA* 104, 13868–13874. doi: 10.1073/pnas.0706627104

Beer, B., and Bender, A. (2015). Causal inferences about others' behavior among the Wampar, Papua New Guinea—and why they are hard to elicit. *Front. Psychol.* 6:128, 1–14. doi: 10.3389/fpsyg.2015.00128

Beller, S., Bender, A., and Song, J. (2009). Weighing up physical causes: effects of culture, linguistic cues, and content. *J. Cogn. Cult.* 9, 347–365. doi: 10.1163/156770909x12518536414493

Bender, A. (2019). The role of culture and evolution for human cognition. *Top. Cogn. Sci.* doi: 10.1111/tops.12449 [online publication ahead of print].

Bender, A., and Beller, S. (2011). Causal asymmetry across cultures: assigning causal roles in symmetric physical settings. *Front. Psychol. Cult. Psychol.* 2:231. doi: 10.3389/fpsyg.2011.00231

Bender, A., and Beller, S. (2016). Probing the cultural constitution of causal cognition – a research program. *Front. Psychol.* 7:245, 1–6. doi: 10.3389/fpsyg.2016.00245

Bender, A., and Beller, S. (2017). Agents and patients in physical settings: linguistic cues affect the assignment of causality in German and Tongan. *Front. Psychol. Cogn. Sci.* 8:1093. doi: 10.3389/fpsyg.2017.01093

Bender, A., and Beller, S. (2019). The cultural fabric of human causal cognition. *Perspect. Psychol. Sci.* 14, 922–940. doi: 10.1177/1745691619863055

Bender, A., Beller, S., and Medin, D. L. (2017). "Causal cognition and culture" in *The Oxford handbook of causal reasoning*. ed. M. R. Waldmann (New York: Oxford University Press), 717–738.

Blaisdell, A. P., Sawa, K., Leising, K. J., and Waldmann, M. R. (2006). Causal reasoning in rats. *Science* 311, 1020–1022. doi: 10.1126/science.1121872

Bohnemeyer, J., Enfield, N. J., Essegbey, J., and Kita, S. (2010). "The macro-event property: the segmentation of causal chains" in *Event representation in language*. eds. J. Bohnemeyer and E. Pederson (Cambridge: Cambridge University Press), 43–67.

Callanan, M. A., and Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: causal thinking in everyday activity. *Cogn. Dev.* 7, 213–233. doi: 10.1016/0885-2014(92)90012-G

Carey, S. (1996). "Cognitive domains as modes of thought" in *Modes of thought: Explorations in culture and cognition*. eds. D. Olson and N. Torrance (Cambridge: Cambridge University Press), 187–215.

Carey, S. (2009). *The origin of concepts*. Oxford: Oxford University Press.

Choi, I., Dalal, R., Kim-Prieto, C., and Park, H. (2003). Culture and judgement of causal relevance. *J. Pers. Soc. Psychol.* 84, 46–59. doi: 10.1037/0022-3514.84.1.46

Choi, I., Nisbett, R. E., and Norenzayan, A. (1999). Causal attribution across cultures: variation and universality. *Psychol. Bull.* 125, 47–63. doi: 10.1037/0033-2909.125.1.47

Clegg, J. M., and Corriveau, K. H. (2017). Children begin with the same start-up software, but their software updates are cultural. *Behav. Brain Sci.* 40:32. doi: 10.1017/s0140525x17000097

Csibra, G., and Gergely, G. (2009). Natural pedagogy. *Trends Cogn. Sci.* 13, 148–153. doi: 10.1016/j.tics.2009.01.005

D'Andrade, R. G. (1995). *The development of cognitive anthropology*. Cambridge: Cambridge University Press.

Dennett, D. C., and Lambert, E. (2017). Thinking like animals or thinking like colleagues? *Behav. Brain Sci.* 40, 34–35. doi: 10.1017/s0140525x17000127

Dickinson, A., and Balleine, B. W. (2000). "Causal cognition and goal-directed action" in *Vienna series in theoretical biology: The evolution of cognition*. eds. C. Heyes and L. Huber (Cambridge: The MIT Press), 185–204.

Dörner, D. (1996). *The logic of failure*. New York: Basic Books.

Fausey, C. M., and Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. *Psychon. Bull. Rev.* 17, 644–650. doi: 10.3758/PBR.17.5.644

Fausey, C., Long, B., Inamori, A., and Boroditsky, L. (2010). Constructing agency: the role of language. *Front. Psychol.* 1:162. doi: 10.3389/fpsyg.2010.00162

Frazier, B. N., Gelman, S. A., and Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult–child conversation. *Child Dev.* 80, 1592–1611. doi: 10.1111/j.1467-8624.2009.01356.x

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* 111, 3–32. doi: 10.1037/0033-295X.111.1.3

Gopnik, A., Meltzoff, A. N., and Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn*. New York: William Morrow & Co.

Gopnik, A., and Sobel, D. (2000). Detecting blickets: how young children use information about novel causal powers in categorization and induction. *Child Dev.* 71, 1205–1222. doi: 10.1111/1467-8624.00224

Güss, C. D., and Robinson, B. (2014). Predicted causality in decision making: the role of culture. *Front. Psychol.* 5:479. doi: 10.3389/fpsyg.2014.00479

Iliev, R., and Axelrod, R. (2016). Does causality matter more now? Increase in the proportion of causal language in English texts. *Psychol. Sci.* 27, 635–643. doi: 10.1177/0956797616630540

Iliev, R., and ojalehto, B. (2015). Bringing history back to culture: on the missing diachronic component in the research on culture and cognition. *Front. Psychol.* 6:716. doi: 10.3389/fpsyg.2015.00716

Kelley, H. H. (1973). The processes of causal attribution. *Am. Psychol.* 28, 107–128. doi: 10.1037/h0034225

Kempton, W. M. (1986). Two theories of home heat control. *Cogn. Sci.* 10, 75–90. doi: 10.1207/s15516709cog1001_3

Kuhnmünch, G., and Beller, S. (2005). Distinguishing between causes and enabling conditions – through mental models or linguistic cues? *Cogn. Sci.* 29, 1077–1090. doi: 10.1207/s15516709cog0000_39

Kushnir, T., Wellman, H. M., and Gelman, S. A. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition* 107, 1084–1092. doi: 10.1016/j.cognition.2007.10.004

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40, 1–72. doi: 10.1017/s0140525x16001837

Legare, C. H., and Gelman, S. A. (2008). Bewitchment, biology, or both: the co-existence of natural and super-natural explanatory frameworks across development. *Cogn. Sci.* 32, 607–642. doi: 10.1080/03640210802066766

Leslie, A. M. (1994). "ToMM, ToBy, and agency: Core architecture and domain specificity" in *Mapping the mind: Domain specificity in cognition and culture*. eds. L. A. Hirschfeld and S. A. Gelman (Cambridge: Cambridge University Press), 118–148.

Livesey, E. J., Goldwater, M. B., and Colagiuri, B. (2017). Will human-like machines make human-like mistakes? *Behav. Brain Sci.* 40:41. doi: 10.1017/S0140525X1700019X

Loftus, E. F., and Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *J. Verbal Learn. Verbal Behav.* 13, 585–589. doi: 10.1016/S0022-5371(74)80011-3

Lombard, M., and Gärdenfors, P. (2017). Tracking the evolution of causal cognition in humans. *J. Anthropol. Sci.* 95, 1–16. doi: 10.4436/JASS.95006

Lyons, D. E., Young, A. G., and Keil, F. C. (2007). The hidden structure of overimitation. *Proc. Natl. Acad. Sci. USA* 104, 19751–19756. doi: 10.1073/pnas.0704452104

Maddux, W. W., and Yuki, M. (2006). The "ripple effect": cultural differences in perceptions of the consequences of events. *Pers. Soc. Psychol. Bull.* 32, 669–683. doi: 10.1177/0146167205283840

McGuigan, N., Makinson, J., and Whiten, A. (2011). From over-imitation to super-copying: adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. *Br. J. Psychol.* 102, 1–18. doi: 10.1348/000712610X493115

Medin, D. L., and Atran, S. (2004). The native mind: biological categorization and reasoning in development and across cultures. *Psychol. Rev.* 111, 960–983. doi: 10.1037/0033-295x.111.4.960

Michotte, A. E. (1963). *The perception of causality*. London: Methuen [Original published in 1946].

Morris, M. W., and Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *J. Pers. Soc. Psychol.* 67, 949–971. doi: 10.1037/0022-3514.67.6.949

Muentener, P., and Bonawitz, E. B. (2017). "The development of causal reasoning" in *The Oxford handbook of causal reasoning*. ed. M. R. Waldmann (New York: Oxford University Press), 677–698.

Norenzayan, A., and Nisbett, R. E. (2000). Culture and causal cognition. *Curr. Dir. Psychol. Sci.* 9, 132–135. doi: 10.1111/1467-8721.00077

ojalehto, B., Medin, D. L., and García, S. G. (2017a). Conceptualizing agency: Folkpsychological and folkcommunicative perspectives on plants. *Cognition* 162, 103–123. doi: 10.1016/j.cognition.2017.01.023

ojalehto, B., Medin, D. L., and García, S. G. (2017b). Grounding principles for inferring agency: two cultural perspectives. *Cogn. Psychol.* 95, 50–78. doi: 10.1016/j.cogpsych.2017.04.001

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: MIT Press.

Pearl, J., and Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. New York: Basic Books.

Penn, D. C., and Povinelli, D. J. (2007). Causal cognition in human and nonhuman animals: a comparative, critical review. *Annu. Rev. Psychol.* 58, 97–118. doi: 10.1146/annurev.psych.58.110405.085555

Saxe, R., and Carey, S. (2006). The perception of causality in infancy. *Acta Psychol.* 123, 144–165. doi: 10.1016/j.actpsy.2006.05.005

Shanks, D. R., Holyoak, K., and Medin, D. L. (Eds.) (1996). *Causal learning*. San Diego: Academic Press.

Spelke, E. S., and Kinzler, K. D. (2007). Core knowledge. *Dev. Sci.* 10, 89–96. doi: 10.1111/j.1467-7687.2007.00569.x

Sperber, D., Premack, D., and Premack, A. J. (Eds.) (1995). *Causal cognition: A multidisciplinary debate*. Oxford: Clarendon Press.

Stuart-Fox, M. (2015). The origins of causal cognition in early hominins. *Biol. Philos.* 30, 247–266. doi: 10.1007/s10539-014-9462-y

Tessler, M. H., Goodman, N. D., and Frank, M. C. (2017). Avoiding frostbite: it helps to learn from others. *Behav. Brain Sci.* 40, 48–49. doi: 10.1017/s0140525x17000280

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* 28, 675–735. doi: 10.1017/S0140525X05000129

Waldmann, M. R. (Ed.) (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.

Waldmann, M. R., Hagmayer, Y., and Blaisdell, A. P. (2006). Beyond the information given: causal models in learning and reasoning. *Curr. Dir. Psychol. Sci.* 15, 307–311. doi: 10.1111/j.1467-8721.2006.00458.x

Wolff, P., Jeon, G. H., and Li, Y. (2009). Causers in English, Korean, and Chinese and the individuation of events. *Lang. Cogn.* 1, 167–196. doi: 10.1515/LANGCOG.2009.009

Zuberbühler, K. (2000). Causal cognition in a non-human primate: field playback experiments with Diana monkeys. *Cognition* 76, 195–207. doi: 10.1016/S0010-0277(00)00079-2

Conflict of Interest: 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 February 2020
doi: 10.3389/fpsyg.2020.00244

Causal Structure Learning in Continuous Systems

Zachary J. Davis[1]*, *Neil R. Bramley*[2] and *Bob Rehder*[1]

[1] Department of Psychology, New York University, New York, NY, United States, [2] Department of Psychology, The University of Edinburgh, Edinburgh, United Kingdom

Real causal systems are complicated. Despite this, causal learning research has traditionally emphasized how causal relations can be induced on the basis of idealized events, i.e., those that have been mapped to binary variables and abstracted from time. For example, participants may be asked to assess the efficacy of a headache-relief pill on the basis of multiple patients who take the pill (or not) and find their headache relieved (or not). In contrast, the current study examines learning via interactions with continuous dynamic systems, systems that include continuous variables that interact over time (and that can be continuously observed in real time by the learner). To explore such systems, we develop a new framework that represents a causal system as a network of stationary Gauss–Markov ("Ornstein–Uhlenbeck") processes and show how such *OU networks* can express complex dynamic phenomena, such as feedback loops and oscillations. To assess adult's abilities to learn such systems, we conducted an experiment in which participants were asked to identify the causal relationships of a number of OU networks, potentially carrying out multiple, temporally-extended interventions. We compared their judgments to a normative model for learning OU networks as well as a range of alternative and heuristic learning models from the literature. We found that, although participants exhibited substantial learning of such systems, they committed certain systematic errors. These successes and failures were best accounted for by a model that describes people as focusing on pairs of variables, rather than evaluating the evidence with respect to the full space of possible structural models. We argue that our approach provides both a principled framework for exploring the space of dynamic learning environments as well as new algorithmic insights into how people interact successfully with a continuous causal world.

Keywords: causal learning, dynamic systems, computational modeling, intervention, cognitive modeling, resource limitations

OPEN ACCESS

Edited by:
York Hagmayer,
University of Göttingen, Germany

Reviewed by:
Joachim Funke,
Heidelberg University, Germany
Wolfgang Schoppek,
University of Bayreuth, Germany

****Correspondence:***
Zachary J. Davis
zach.davis@nyu.edu

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

Received: *07 October 2019*
Accepted: *03 February 2020*
Published: *20 February 2020*

Citation:
Davis ZJ, Bramley NR and Rehder B
(2020) Causal Structure Learning in
Continuous Systems.
Front. Psychol. 11:244.
doi: 10.3389/fpsyg.2020.00244

INTRODUCTION

We live and act in a messy world. Scientists' best models of real-world causal processes typically involve not just stochasticity, but real-valued variables, complex functional forms, delays, dose-dependence, and feedback leading to rich and often non-linear emergent dynamics (Cartwright, 2004; Strevens, 2013; Sloman and Lagnado, 2015). It follows that learning successfully in natural settings depends on accommodating these factors. Cognitive psychologists have explored many of these dimensions of complexity in isolation (e.g., *stochasticity:* Waldmann and Holyoak, 1992; Bramley et al., 2017a; Rothe et al., 2018; *interventions:* Sloman and Lagnado, 2005; Waldmann and Hagmayer, 2005; Bramley et al., 2015; Coenen et al., 2015; *time:* Buehner and May, 2003; Lagnado and Sloman, 2006; Rottman and Keil, 2012; Bramley et al., 2018; and *continuous variables:* Pacer and Griffiths, 2011). However, we argue these components generally can not be isolated in realistic learning settings, meaning a deeper understanding of human causal cognition will require a new framework that naturally accommodates inference from interventions in continuous dynamic settings.

As an everyday example of a time-sensitive, dose-dependent causal relationship, consider the complexities involved in consuming alcohol. It is common for drinkers to adjust their consumption based on their recognition that higher doses affect inhibition or mental clarity, that will in turn have other downstream effects on quality of conversation or willingness to sing karaoke. The effects of alcohol consumption differ widely in quality and quantity depending on dosage and time delays. A small glass of wine with dinner will likely have little effect on mental clarity whereas a few shots will have a stronger effect. Further complicating the learning problem, these effects of alcohol do not come instantaneously but are rather delayed and distributed in time. Worse still, there can be complex temporal dynamics, such as the feedback loop between lowered inhibition and increased alcohol consumption, and innumerable contributing factors, such as diet or amount of sleep, that modulate alcohol's effect. Thus, in settings like this, the learning problem is non-discrete (how much alcohol did I drink) and extended in time (when did I drink it), produces evidence that is naturally time ordered (how you feel over the preceding and subsequent hours), and involves complicated dynamics (e.g., feedback loops). In the current paper, we study human learning through real-time interactions with causal systems made up of continuous valued variables. We see this setting as capturing the richness of real world causal learning, while remaining simple and principled enough to allow for a novel formal analysis.

The structure of the paper is as follows. First, we summarize relevant past work on causal structure inference from interventions, temporal information, and different representations of functional form. Next, we lay out our new formalism for inference of causal structure between continuous variables. We then report on an experiment, in which participants interact with causal systems represented by sliders on the computer screen. We provide an exploratory analysis of the interventional strategies we observed in the experiment before analyzing structure learning through the lens of a normative Bayesian inference model and a range of heuristic and approximate alternatives, finding evidence that people focus sequentially on individual connections rather than attempting to learn across the full space of possible causal models at once. Finally, we discuss new opportunities provided by the formalism introduced in this paper, including future questions in causal cognition as well as applications to other areas, such as dynamic control.

Past Research

Probabilistic Causation Over Discrete Events

Research in causal cognition has generally aligned itself with the philosophical tradition of probabilistic causation, which defines a causal relationship as one where a cause changes the probability of its effect (Hitchcock, 2018). This definition implicitly operates over particular representations: discrete states, such as events or facts that have some probability of occurring or being true. Because of this, experimental work in causal cognition has primarily focused on causal relationships between discrete valued (often binary) variables (e.g., Sloman, 2005; Krynski and Tenenbaum, 2007; Ali et al., 2011; Fernbach and Erb, 2013; Hayes et al., 2014; Rehder, 2014; Rothe et al., 2018). These are typically presented in contexts in which temporal information is either unavailable or abstracted away so that cases can be summarized in a contingency table. See **Figure 1** for a simple example in which (A) continuous data is (B) snapshotted in time, in order to (C) dichotomize and create counts of contingencies and ultimately abstracted into a probabilistic causal relationship. This approach is very common in part because there is a well-established mathematical framework—*Bayesian networks*—for efficiently encoding joint distributions of sets of variables in the form of networks of probabilistic contingencies (Pearl, 2009; Barber, 2012).

While the probabilistic contingencies paradigm has been fruitful for exploring many aspects of causal cognition, we are interested in other settings. As mentioned, we believe that many real life systems may not lend themselves to discretization, nor involve much independent and identically distributed data with no temporal information. Instead, people are often have access to autocorrelated, time-dependent, continuous information and we are interested in they how represent and draw inferences on the basis of this information.

Learning

A prominent question in causal cognition is how people learn causal relationships from contingency data, such as that presented in **Figure 1C**. Although the literature shows that humans are often quite adept causal learners (Cheng, 1997; Griffiths and Tenenbaum, 2005; Lu et al., 2008) there are a number of important exceptions. One is that updates to beliefs about causal structure on the basis of new information are often made narrowly rather than globally. That is, in ways that do not compare the evidential fit across all variables taken together. To model this, Fernbach and Sloman introduced a *Local Computations* (LC) model, which posits that people focus on "evidence for individual causal relations rather than evidence for fully specified causal structures" (Fernbach and Sloman, 2009, p. 680). By ignoring the possible influences of other causes, their model captures a strong empirical tendency for human learners to exhibit order effects and overconnect their causal hypotheses (also see Taylor and Ahn, 2012). Bramley et al. (2017a) extended this finding, finding evidence suggesting that people consider local changes that modify their previously favored hypothesis. Together, these studies suggest that people use a local updating strategy, testing and evaluating individual causal links rather than updating a posterior distribution over the global model space. We

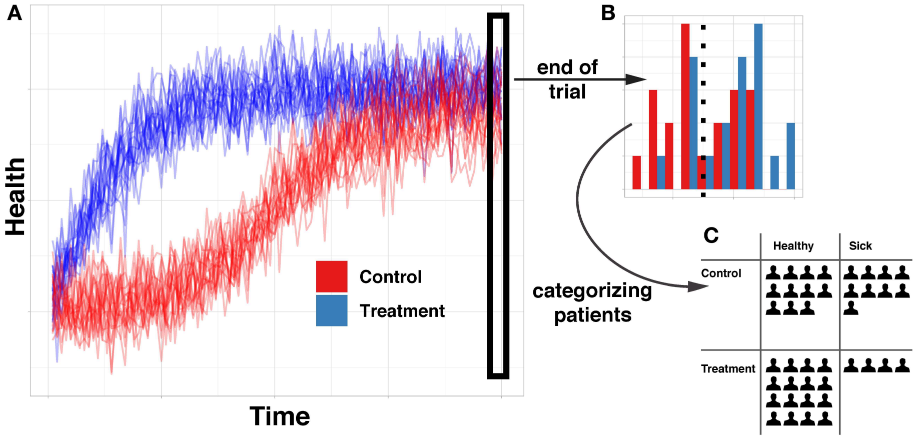


FIGURE 1 | Illustration of abstraction from full timeseries data to probabilistic contingency. **(A)** is a full time course of the health of 40 simulated patients throughout the course of a classic randomized controlled trial. **(B)** demonstrates the type of information available when only evaluating the health of patients at the end of the trial. **(C)** demonstrates the type of information available when categorizing patients into "sick" and "healthy" groups, rather than maintaining full continuous information.

ask whether this tendency toward local learning extends to the continuous dynamic systems that are under study here.

Learning via Interventions

As well as capturing probabilistic relationships, Bayesian networks can be used to reason about, and from, idealized manipulations of causal systems, or "interventions" (Pearl, 2009). Bayesian networks, at their core, deal with *independence*, not dependence, relations. Because of this, if a cognizer passively observes some variables but cannot observe the temporal direction of their influences (i.e., perhaps they influence one another too quickly to see) they can be equally consistent with multiple causal hypotheses. For example, the common cause $X \leftarrow Y \rightarrow Z$ and chain $X \rightarrow Y \rightarrow Z$ are "Markov equivalent" because, in both networks, X and Z are independent conditional on Y. However, crucially, Markov equivalent networks do not have identical data distributions under intervention. In the example of Markov equivalent networks given above, intervening to set Y to some value y as denoted with Pearl's (2009) "Do()" operator, would change the distribution for X under the common cause—i.e., $P(X) \neq P(X|\text{Do}[Y = y])$ for at least some y—but would not affect the distribution for X for the chain—i.e., $P(X) = P(X|\text{Do}[Y = y])$ for any y.

It has been shown that people are able to learn successfully from interventions, and are often moderately efficient in their intervention selection according to information–optimal norms (Steyvers et al., 2003; Sloman and Lagnado, 2005; Waldmann and Hagmayer, 2005; Coenen et al., 2015; Bramley et al., 2017a). However, participants in these studies also typically exhibited biases indicative of the influence of cognitive constraints. For example, Coenen et al. (2015) found that, when deciding between two potential causal networks, people appeared to follow a heuristic of intervening on the node with the most downstream causal links (averaged across the candidate networks) rather than intervening to maximally distinguish between the two. Use of this heuristic was more common when intervening under time pressure. Bramley et al. (2017a) tested people's learning in a broader hypothesis space encompassing all possible 3 and 4 variable network structures. They found that people made interventions that appeared to target uncertainty about a specific individual link, node or confirm a single hypothesis, rather than those effective at reducing their uncertainty "globally" over all possible causal networks. Here we assess the efficacy of learners' interventions on continuous dynamic systems for which variables are potentially manipulated through a range of magnitudes over an extended period of time.

Time

Time has long been seen as a powerful cue for causation (Hume, 1959), especially with regards to identifying causal direction. People rule out backwards causation, assuming that effects cannot precede causes (Burns and McCormack, 2009; Greville and Buehner, 2010; Bramley et al., 2014). Work in the cognitive sciences on the use of time in causal judgments has focused on point events separated by delays—that is, events like explosions and collisions that occur at particular times but with negligible duration (Shanks et al., 1989; Griffiths, 2004; Lagnado and Sloman, 2006; Pacer and Griffiths, 2012; McCormack et al., 2015). From this line of work, we have learned more than just that temporal order is relevant for causal direction. The actual temporal dynamics of causal systems affect judgments, for example shorter and more reliable delays between cause and effect are more readily seen as causal (Greville and Buehner, 2010).

In a systematic study of people's use of temporal dynamics to learn causal structure, Bramley et al. (2017b, 2018) combined interventions and time to investigate people's learning of causal structure between components that exhibited occasional (punctate) events that could also be brought about by interventions. They found that people are sensitive to expected delays, especially when they also expect the true delays to be reliable, and are judicious and systematic in their use of interventions. While these studies have been valuable in demonstrating that people are sensitive to the temporal characteristics of causal systems, many everyday systems—such as economies, ecosystems, or social groups—are more naturally described as extended shifting influences than point events. We thus see the current study as extending the analysis of time's role in causal cognition to explore these inherently continuous settings.

Continuous Variables

As discussed above, many natural scenarios involve continuous valued variables and causal influences that are typically extended in time rather than punctate. Given the ubiquity of such systems, continuous variables have received surprisingly little attention in the study of causal cognition. In a reanalysis of data from Marsh and Ahn (2009) and a novel experiment, Pacer and Griffiths (2011) showed that people are capable of learning individual cause-effect relationships between continuous variables. Soo and Rottman (2018) investigated causal relations in non-stationary time series, i.e., those where long term trends affect the average values of the variables in ways that obscure and complicate the causal relations between those variables. They proposed three ways that the variables could be represented before assessing their relationships: (1) state values, (2) difference scores, and (3) trinarized difference scores (positive, negative, or zero). In their task, causal strength judgments were best explained by the correlation between the direction of *changes* in variables' values from one time point to the next, rather than direct correlation between the variables.

Complex Problem Solving

This project connects to the literature on complex problem solving (Berry and Broadbent, 1984)—also sometimes called complex dynamic control (Osman, 2010). This line of work explores goal-directed behavior in dynamic environments, typically with a structure that is hidden and initially unknown to participants. In particular, we follow Funke (2001) in studying minimal complex systems (MICS) that change dynamically in response to participants' actions and their hidden structure, but are not so complex as to prohibit formal analysis. MICS have been used as psychometric measurement tools, having been shown to provide individually stable and reliable predictors of real-world achievement (Greiff et al., 2013). This suggests that MICS tap into fairly foundational cognitive abilities.

Research on complex problem solving has begun to unpack the key features of such MICS, and of the cognitive strategies recruited by participants that determine performance. For example, when participants have narrow goals in a new environment, they learn less about its overall structure (Vollmeyer et al., 1996), a finding consistent with proposals that monitoring goals induces cognitive demands (Sweller, 1988). They are also less likely to engage in systematic strategies that can aid learning, such as the Vary One Thing At a Time (VOTAT, see Kuhn and Brannock, 1977; Tschirgi, 1980) or PULSE strategy (Schoppek and Fischer, 2017). Other work has identified a number of high level behavioral features, such as time on task, number of interventions made, or strategies, that predict likelihood of success (Greiff et al., 2016; Schoppek and Fischer, 2017; Stadler et al., 2019).

We build on previous work in the CPS literature in a number of ways. For one, whereas tasks in the CPS literature are typically self-paced, we are unusual (but not unique, see Brehmer and Allard, 1991; Schoppek and Fischer, 2017) in studying time-continuous systems. We take the task of reacting to dynamics as they unfold in real time to be reflective of real world dynamic control scenarios. More fundamentally, the research area's focus on predicting success in *control* has left a gap in our understanding of what exactly participants are *learning* as they interact with dynamic systems. The current work extends on this line of enquiry by providing a close model-based analysis of participants' actions and learning.

In sum, our approach here is novel in two key respects. First, we study a setting that, like reality, is continuous in terms of both time and state space. This allows us to study learning in the context of causal systems that give rise to non-linear emergent dynamics through the lens of a sophisticated normative and heuristic model comparison. Second, we explore an interactive setting in which participants intervene on the system of interest in complex, extended ways, rather than merely passively observing its behavior or setting states across discrete trials, again mapping more onto real world actions than the idealized interventions studied in much of the existing causal learning literature.

THE TASK

We chose a simple and intuitive structure learning task interface that allows for learners to use their mouse to interact with the variables in a system represented by a set of moving sliders on the computer screen. A depiction of how the sliders were presented is shown in **Figure 4**. Participants could observe the evolving sequence of variable values but also move and hold the variables (one at a time) at positions of their choice by using the mouse. As mentioned, this environment allows us to test learning of causal systems with continuous valued variables and feedback dynamics. It also allows us to assess learning via interventions that are both extended over time (learners choose how long to intervene) and non-stationary (learners might "hold" the variable in a particular position or "wiggle" it up and down).

CONTINUOUS CAUSALITY IN TIME

This section presents a formalism for modeling causal systems that relate continuous variables in time. To define a generative model for such systems, we first introduce the notion of an Ornstein–Uhlenbeck (OU) process and then define how multiple OU processes can be interrelated so as to form an interacting causal system. We then describe normative inference within this model class on the basis of both observational and interventional data.

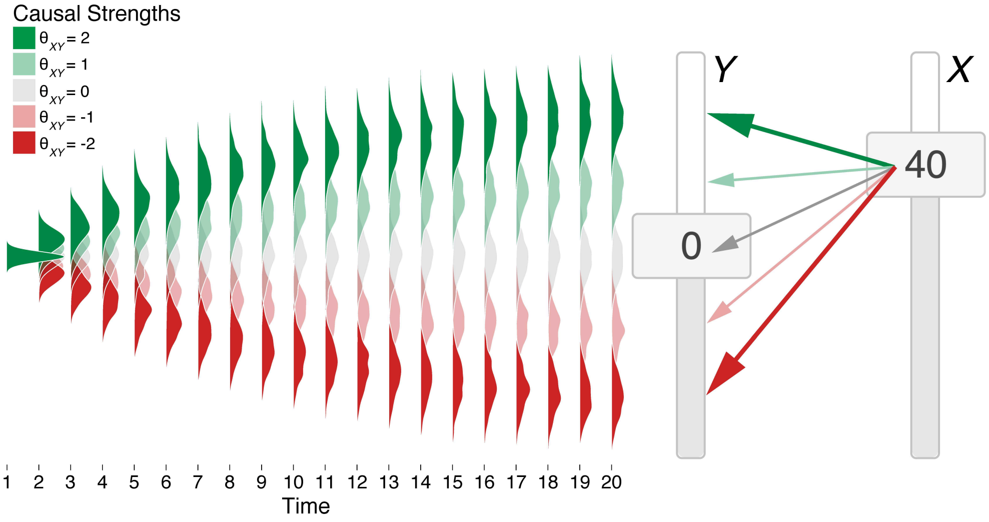


FIGURE 2 | Visualization of the impact of a single cause (slider *X*) on a single effect (slider *Y*) in an OU network with different causal strengths. Slider *X* is held to a value of 40 for 20 timepoints, leading slider *Y* to unfold over time to different values depending on the causal strength. Probability distributions are smoothed averages of 100 runs of the network given different causal "strengths" θ_{XY} (colored shading) where ω = 0.1 and σ = 5.

Generative Model

The Ornstein–Uhlenbeck Process

An Ornstein–Uhlenbeck (OU) process is a stationary Gauss-Markov process that reverts to a stable mean (Uhlenbeck and Ornstein, 1930). It can be conceptualized as Brownian motion with the addition of a corrective force that biases the process's expected value toward the mean of the distribution. The magnitude of that force increases as a function of the distance been that mean and the process's current state. Formally, Δv_i^t—the change in variable *i* from time *t* to $t + 1$—is defined as

$$P(\Delta v_i^t|\omega, \mu_i, v_i^t, \sigma) = \omega[\mu_i - v_i^t] + N(0, \sigma) \tag{1}$$

where v_i^t is the value of *i* at time *t*, μ_i is the mean of the process for variable *i*, σ is its variance, and ω is a parameter > 0 that determines how sharply the process reverts to the mean[1]. μ_i is also referred to as the process's *attractor state* because it is the value to which the process will revert to at asymptote. See **Figure 3A** for an example of an OU process with an attractor state of 0.

OU Processes and Causality

This definition can be generalized to accommodate OU processes with non-stationary means. In particular, we take the step of assuming that the attractor state μ for a variable is determined by some function of the most recent values of its cause(s). When a variable has no causes we model its attractor state as being 0.

The single cause case

For a variable *i* with a single cause *j* this function is simply,

$$\mu_i^{t+1} = f(v_j^t) \tag{2}$$

where v_j^t is the value of *j* at time *t*. As *j* changes over time, so too does the output of $f(v_j^t)$, which serves as the new attractor state of variable *i* at the next timepoint. For simplicity, here we assume that $f(v_j^t)$ is linear. Thus, the change in *i* at the next timestep (Δv_i^t) is

$$P(\Delta v_i^t|v_i^t, v_j^t, \omega, \sigma, \theta_{ji}) = \omega[\theta_{ji} \cdot v_j^t - v_i^t] + N(0, \sigma) \tag{3}$$

where $\theta_{ji} \in (-\infty, \infty)$ is a multiplier (or "strength") mapping the value of the cause *j* to the attractor state of effect *i*. **Figure 2** presents how a variable *Y* changes as a function of its cause *X* for a number of different values of θ_{XY}. We assume Δt of 100 ms (i.e., between *t* and $t + 1$) and that ω and σ remain constant, although these assumptions can be loosened (see Lacko, 2012).

The multiple cause case

In general, a variable may have more than one cause. Although there are a variety of ways in which multiple causal influences might combine (cf. Griffiths and Tenenbaum, 2009; Pacer and Griffiths, 2011), here we simply assume that causes have an additive influence on an effects' attractor state, such that

$$P(\Delta v_i^t|v^t, \omega, \sigma, \Theta) = \omega\Big[\Big[\sum_j \theta_{ji} \cdot v_j^t\Big] - v_i^t\Big] + N(0, \sigma) \tag{4}$$

where *j* now ranges over all causes of variable *i* and Θ is a square matrix such that $\theta_{ji} \in \Theta$ is the strength of the causal

[1] Throughout this work we use subscripts to denote variables and superscripts to denote time. Note that whereas v_i^t is the value of *i* at time *t*, v_i is the value of *i* at all timesteps, v^t is the value of all variables at time *t*, and *v* is the value of all variables at all times.

relationship from j to i[2]. Simply put, the mean that variable i reverts to is assumed to be a sum of the values of its causes, each first multiplied by their respective θs.

A collection of connected OU processes, which we call an *OU network*, defines causal relationships for all directed relations between variables and unrolls these effects over time. For example, for a system consisting of variables X, Y, and Z, Θ specifies the strengths of the six potential inter-variable causal relationships: $X \rightarrow Y$, $Y \rightarrow X$, $X \rightarrow Z$, $Z \rightarrow X$, $Y \rightarrow Z$, and $Z \rightarrow Y$. Note that non-relationships are specified in this scheme by setting θ_{ji} to zero. At each timestep, Equation (4) is used to determine v_X^{t+1}, v_Y^{t+1}, and v_Z^{t+1} as function of their previous values v_X^t, v_Y^t, and v_Z^t. For display purposes, it is sometimes necessary to constrain v to be between some range. This is done by setting all v^{t+1} that fall outside of the range to their nearest value in the range. The clock then moves forward and the process repeats.

OU networks have some intuitively appealing features of continuously varying causal relationships. **Figure 3** demonstrates some of the dynamics that emerge from causal systems simply by varying the θs. Whereas, a positive θ_{XY} results in the value of Y following some positive multiple of the value of X (**Figure 3B**), a negative θ_{XY} means that a decrease in X drives up the value of Y (e.g., decreasing interest rates is generally thought to increase inflation, **Figure 3C**). Feedback loops are naturally represented with non-zero values of θ_{XY} and θ_{YX}. A positive feedback loop results if the θs are of the same sign and have an average magnitude >1 (**Figure 3D**) whereas a negative feedback loop results if they are <1 (**Figure 3E**). Oscillations can be implemented with θs of mismatched signs (such as 5 and -5, **Figure 3F**). Such feedback loops can be implemented between pairs of variables or as part of a cyclic causal structure with potentially many variables. Combining feedback loops and cycles and including asymmetrical forms can lead to even more complex dynamics (e.g., **Figure 3H**). We invite the reader to build their own network and observe the dynamics at https://zach-davis.github.io/html/ctcv/demo_ctcv.html. Note that while the discussed examples cover two or three variables, the OU networks framework generalizes to any number of variables.

Inference

We follow Griffiths and Tenenbaum (2005) in modeling people's learning of causal graphs as inverting the generative model. What must be inferred is the causal structure most likely responsible for producing all variable values at all timepoints—v—under interventions.

Note that to accommodate interventions, we adopt Pearl's (2009) notion of graph surgery. If variable i is manipulated at time t, the likelihood that v_i^t has its observed value is 1 (i.e., is independent of i's previous value or the value of its causes). We define ι_i^t as an indicator variable that is true if variable i is intervened on at t and false otherwise.

The Single Cause Case

Consider the inference problem in which the goal is to determine whether variable j causes variable i, and if so, the sign of that causal relationship. That is, assume a hypothesis space L with three hypotheses. One is that θ_{ji} is >0, a causal relationship we refer to as a *regular connection*. A second is that θ_{ji} is <0, referred to as an *inverse connection*. Finally, $\theta_{ji} = 0$ denotes that j has no impact on i. Assume that i has no other potential causes.

Computing the posterior probability of a causal hypothesis $l_k \in L$ involves computing, for each timepoint t, the likelihood of the observed change in i (Δv_i^t) given the previous values of i and j (v_i^t and v_j^t), a value of θ_{ji} corresponding to the hypothesis, the endogenous system parameters ω and σ, and any intervention that may have occurred on i (ι_i^t). If the learner did not intervene on i at t, this likelihood is given by Equation (3). If they have, it is 1. The product of these likelihoods over all timepoints is proportional to the posterior probability of l_k.

$$P(l_k|v_i, v_j; \iota_i) \propto \prod_t \int_\omega \int_{\theta_{ji}} \int_\sigma P(\Delta v_i^t | v_i^t, v_j^t, \omega, \sigma, \theta_{ji}; \iota_i^t)$$
$$P(\theta_{ji}|l_i) P(l_k) P(\omega) P(\sigma) \mathrm{d}\sigma\, \mathrm{d}\theta_{ji} \mathrm{d}\omega \tag{5}$$

$P(\omega)$ and $P(\sigma)$ represents the learner's prior beliefs about ω and σ. $P(\theta_{ji}|l_k)$ represents the priors over θ_{ji} corresponding to hypothesis l_k. For example, if l_k corresponds to a regular connection, $P(\theta_{ji}|l_k)$ would be 0 for non-positive values of θ_{ji}. For positive values, it would reflect learner's priors over θ_{ji} for regular connections (later we describe how these priors can be estimated in our experiment on the basis of an instructional phase that precedes the causal learning task). Applying Equation (5) to each causal hypothesis and then normalizing yields the posterior over the three hypotheses in L.

A complication arises if variable values v are truncated between some range of values (in our task $v \in [-100, 100]$). In the case where v_i^t equals the maximum truncated value, the likelihood is the mass of the likelihood distribution above the range of values. For the minimum truncated value the likelihood is the mass of the likelihood distribution below the range of values.

The Multiple Cause Case

This procedure for evaluating a single potential causal relationship generalizes to determining the structure of an entire OU network. Consider a hypothesis space G as consisting of *graphs* where each graph defines, for every potential causal relationship, whether it is positive, inverse, or zero. For a system with n variables G would contain 3^{2n} distinct causal hypotheses; for our example system with variables X, Y, and Z, G contains 729 graphs. The posterior probability of a graph $g_k \in G$ involves computing for each variable i and timepoint t, the likelihood of the observed Δv_i^t given the θs defined by g_k and the state of the system's variables at t (Equation 4), taking into account the possibility of an intervention on i at t (ι_i^t):

$$P(g_k|v; \iota) \propto \prod_{i=1}^{N} \prod_t \int_\omega \int_\theta \int_\sigma P(\Delta v_i^t | v^t, \omega, \sigma, \theta; \iota_i^t)$$
$$P(\theta|g_k) P(g_k) P(\omega) P(\sigma) \mathrm{d}\sigma\, \mathrm{d}\theta\, \mathrm{d}\omega \tag{6}$$

[2] Although the OU formalism allows it, throughout this work we ignore the possibility of self-cycles, that is, instances in which variables is a cause of itself. That is, we assume, $\Theta_{ii} = 0$.

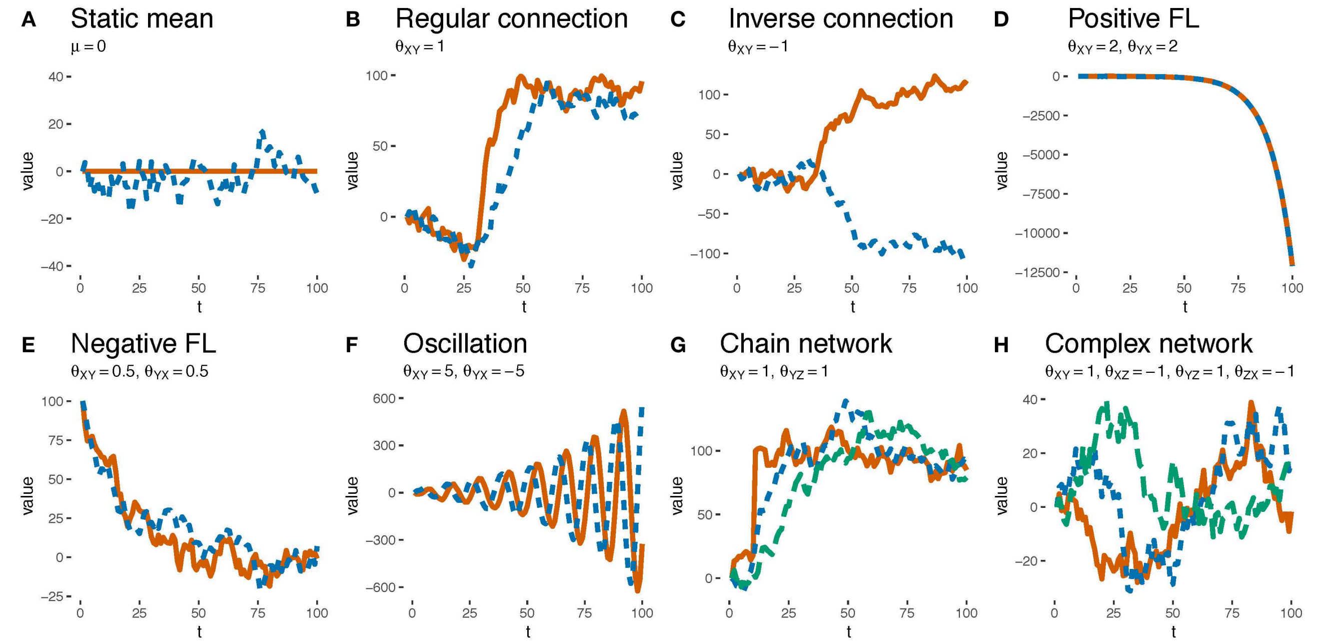


FIGURE 3 | Examples of the dynamical phenomena resultant from varying θ weights. Solid red, dotted blue, and dashed green lines depict the values of variables X, Y, and Z, respectively. **(A)** A system with a single variable Y whose distribution mean is stationary at 0 (i.e., $\mu = 0$). **(B)** A system with variables X and Y and a θ weight from X and Y of 1 (i.e., $\theta_{XY} = 1$). $\mu_X = 0$ for first 30 timepoints and then $\mu_X = 100$ for next 70. The value of Y tracks the value of X. **(C)** The same as **(B)** except that X and Y are negatively related ($\theta_{XY} = -1$). The value of Y tracks but has the opposite sign of X. **(D)** A system in which X and Y are reciprocally related via θ weights that are >1 (i.e., $\theta_{XY} = \theta_{YX} = 2$). Because the values of X and Y grow so large they are indistinguishable in the plot. **(E)** The same as **(D)** except that X and Y, which have an initial value of 100, are reciprocally related via θ weights that are <1 ($\theta_{XY} = \theta_{YX} =$ 0.5). The values of X and Y eventually fluctuate around 0. **(F)** The same as **(D)** except that the reciprocal θs are large and of opposite sign (i.e., $\theta_{XY} = 5, \theta_{YX} = -5$). The values of X and Y oscillate. **(G)** A system with three variables whose θ weights form a causal chain, $\theta_{XY} = \theta_{YZ} = 1$. μ_X= 0 for 10 timepoints but then is set to 100 via an intervention. Note that changes in Y precede changes in Z. **(H)** Timeseries of actual data observed by participant 10 on trial 10, generated by a complex system with three variables and four non-zero θs. All variables were initialized at 0 and there were no interventions.

EXPERIMENT: CAUSAL STRUCTURE LEARNING

To test people's ability to learn causal structure between continuous variables in continuous time, we conducted an experiment in which participants freely interact with sliders governed by an OU network with hidden causal structure. Their goal was to intervene on the system in order to discover the hidden causal structure.

Method

Participants

Thirty participants (13 female, age M = 37.5, SD = 10.6) were recruited from Amazon Mechanical Turk using psiTurk (Crump et al., 2013; Gureckis et al., 2016). They were paid \$4 for ∼30 min. In a post-test questionnaire, on a ten point scale participants found the task engaging (M = 7.9, SD = 2.2) and not particularly difficult (M = 3.9, SD = 2.6). All procedures were approved by the Institutional Review Board of New York University (IRB-FY2016-231).

Materials

Each of the three variables was represented by a vertical slider that moved by itself according to the underlying OU network but which could also be manipulated by clicking and dragging anywhere on the slider, overriding the state it would otherwise have taken (see **Figure 4**)[3]. A timer was presented at the top of the screen. Participants responded using six additional sliders presented beneath the trial window, one for each potential causal

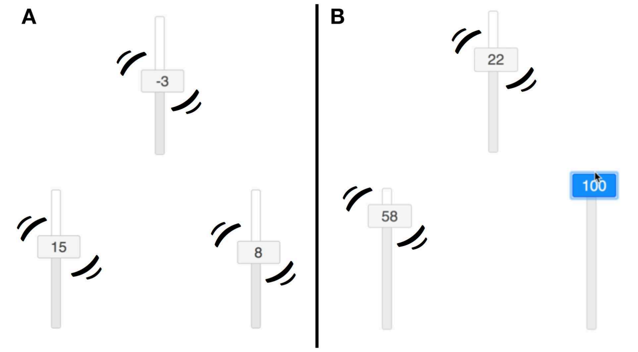


FIGURE 4 | Sliders used by participants. **(A)** Shows that the sliders all jitter if no interventions are made. **(B)** Shows that the sliders do not jitter if intervened on.

[3] See https://zach-davis.github.io/publication/cvct/ for a demo.

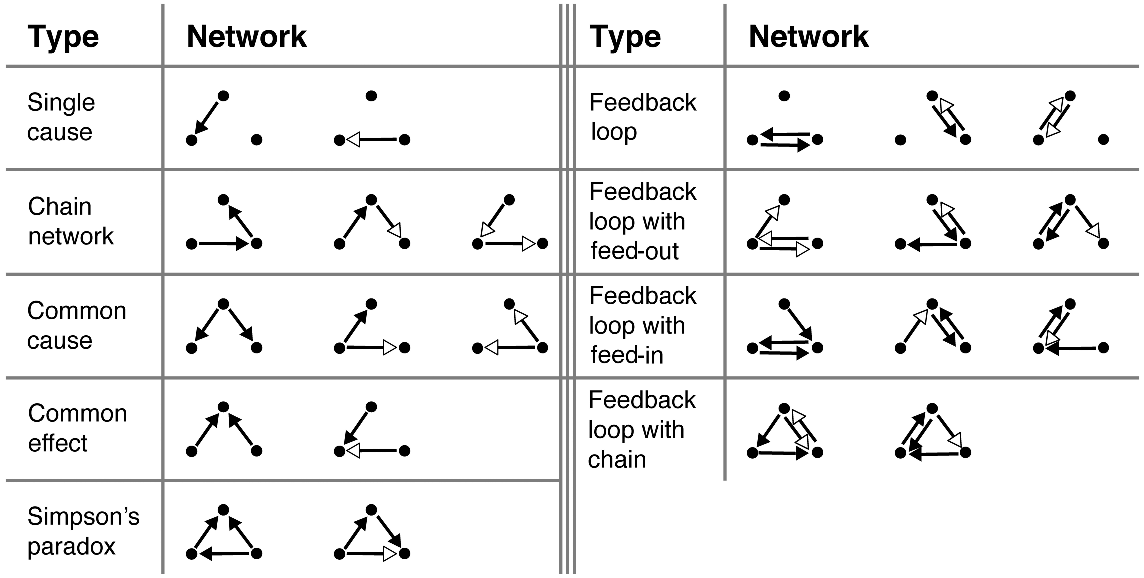


FIGURE 5 | All 23 structures participants were tasked with learning. Black arrowheads signify "regular" connections ($\theta = 1$), white arrowheads signify "inverse" connections ($\theta = -1$).

relations. Responses were constrained to be one of three options: "Inverted," "None," or "Regular," corresponding to $\theta < 0$, no relationship ($\theta = 0$), and $\theta > 0$, respectively. Participants were pre-trained on these terms in the instructions. The sliders were constrained to be between -100 and 100, and the buttons on the slider presented a rounded integer value in addition to moving up and down.

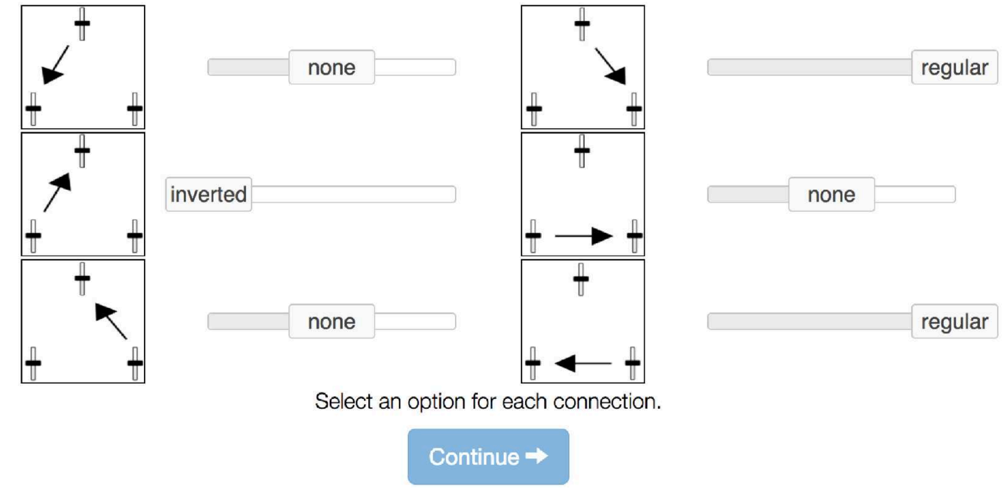
Stimuli and Design

The 23 causal graphs shown in **Figure 5** were selected for testing on the basis of a number of criteria. They were roughly balanced in the number of positive and negative links and the number of links between each of the variables. More qualitatively, we tried to select networks that would be interesting a priori. This includes many of the classic causal graphs, such as chain networks, common causes, and common effects, but also less-studied graphs, such as those with feedback loops. The experiment always began with two practice trials that were excluded from all analyses. These were always the two *Single cause* networks (**Figure 5**, top left). This was followed by 23 test trials, one for each of the networks in **Figure 5** presented in random order. The OU parameters used during training and the test were $\omega = 0.1$ and $\sigma = 5$. The true θs were either 1 (for regular connections), 0 (no connection), or -1 (for inverse connections).

Procedure

To familiarize them with the interface, participants were required to first watch four videos of an agent interacting with example causal networks. These videos informed participants of the underlying causal structure and demonstrated an agent interacting with the system. To minimize biasing participants toward any particular intervention strategy, the videos displayed a variety of basic movements, including wobbling the intervened on variable, holding a variable at a constant level, and holding a variable at a limit value (e.g., 100) by moving its slider to one end of the scale. The four example causal networks included (1) no causal connections, (2) a single regular ($\theta = 1$) connection, (3) a single inverse ($\theta = -1$) connection, and (4) two connections forming a causal chain in which one link was regular and one was inverse. To ensure that they understood the task, participants were required to pass a five question comprehension check before starting. If a participant responded incorrectly to any of the five questions they were permitted to retake the quiz until they responded correctly to all five questions. This was designed to ensure that they learned: the duration of each trial, the difference between a regular and inverted connection, that there can be more than one connection per network, and that they must provide a response for all six possible connections.

In the main task that followed, participants completed 25 trials lasting 45 s each. The first two of these involved a single regular and single inverse connection that, unknown to participants, we considered practice trials to familiarize them with the interface and excluded from all analyses. A trial was initiated by pressing the "Start" button, whereupon the sliders started moving with values updating every 100 ms. Perceptually, they would appear to "jitter" according to the noise associated with the underlying OU network plus move systematically according to the unknown causal relationships. At any time, participants were free to intervene on any variable by clicking, holding, or dragging the requisite slider. While it was pressed down, the position of the mouse determined the value of the variable. Once it was released the variable would continue from that point according to the OU network. Participants were free to make (and revise) their judgments at any point after initiating a trial but were required



none

regular

inverted

none

none

regular

Select an option for each connection.

FIGURE 6 | Judgment options for participants. Participants were presented with a ternary choice between "inverted," "none," and "regular".

to enter a judgement for all six causal relations by the end of the trial (see **Figure 6**). No feedback was provided at any point. After completing the 25 trials, participants completed a brief post-test questionnaire reporting their age, gender, engagement and subjective difficulty as well as any comments.

Results

Participants were substantially above chance (0.33) in correctly classifying causal links into one of the three response categories ($M = 0.82$, $SD = 0.22$), $t_{(29)} = 17.48, p < 0.001$. They were slightly more successful in identifying regular causal links ($M = 0.92$, $SD = 0.12$) than inverse causal links ($M = 0.90$, $SD = 0.13$), $t_{(29)} = 2.12, p = 0.04$. Participants also correctly classified a higher proportion of causal relationships as the trials progressed, as demonstrated by a simple linear regression of accuracy on trial number, $t_{(21)} = 2.91, p = 0.008$, although this relationship was modest with participants being 0.25% more likely to correctly identify a link for each new trial.

In identifying overall causal networks (correctly identifying all six of the possible directional causal relationships), participants were also well above chance ($3^{-6} = 0.0014$), ($M = 0.44$, $SD = 0.22$), $t_{(29)} = 10.81$, p < 0.001. The probability of selecting the correct network was 0.79, 0.60, 25, and 0.07 for networks with 1, 2, 3, and 4 causal links, respectively. Accuracy varied sharply with the complexity of model as shown by a repeated measures ANOVA, $F_{(3,84)} = 74.0$, p < 0.001. Note that participants' responses did not reflect a preference toward simpler models, as they marked slightly over half of the possible connections ($M = 0.52$, $SD = 0.13$), which was greater than the true proportion of connections in the test networks (0.39), $t_{(29)} = 5.62$, p < 0.001. See the Supplementary Material for results for all tested networks.

Errors

While participants were generally well above chance in identifying causal relationships, there was some systematicity to their errors. In particular, these errors closely followed the qualitative predictions of Fernbach and Sloman (2009) local computations (LC) model. The first qualitative prediction is an over-abundance of causal links. Eighty-two percent ($SD = 0.17$) of the errors that participants made involved adding causal links that didn't exist, significantly greater than chance[4] (0.59); $t_{(29)} = 7.33$, $p < 0.001$. The second qualitative prediction of the LC model as defined in this paper is an inability to distinguish between direct and indirect causes (e.g., in the network $X \rightarrow Y \rightarrow Z$, incorrectly also judging $X \rightarrow Z$). While in general participants correctly classified 82% of the causal links, they were far more likely to erroneously add a direct link between two variables when in fact the relationship between those variables was mediated by a third variable, with below chance (0.33) accuracy on those potential links ($M = 0.16, SD = 0.21$); $t_{(29)} = -4.48, p < 0.001$.

Figure 7 shows participant judgments for three classic causal structures in causal cognition: common cause, common effect, and chain networks. It shows that participants were quite good at detecting any causal relationship in a network that existed between two variables. In the figure, these results correspond to the blue bars, which indicate that they correctly classified a regular connection as regular (as mentioned, participants were also good as classifying inverse connections as inverse). **Figure 7** also shows that participants were often good at classifying absent connections as absent (the gray bars) with one important exception: in the chain network $Y \rightarrow Z \rightarrow X$ the relationship between Y and X was judged to be nearly as causal as $Y \rightarrow Z$ and $Z \rightarrow X$. That is, they failed to appreciate that the (apparent) relationship between Y and X was in fact mediated by Z. These patterns held for the other instances of the common cause, common effect, and chain networks defined in **Figure 7**. Moreover, we found that, for any of the more complex networks

[4] For the structures used in this experiment, a hypothetical participant who responded "inverse," "none," and "positive" with equal probability would erroneously add a causal link 59% of the time.

in **Figure 7**, participants had a strong tendency to infer a direct causal relationship between two variables whenever those variables were in fact mediated by the third variable. **Figure S1** presents how causal links were classified for all 23 networks.

Interventions

To achieve this level of performance, participants made heavy use of interventions. We define a single intervention as beginning when a participant clicked on a variable's slider and ending when the mouse was released. The average number of interventions made on a single trial was 4.94 ($SD = 2.46$). However, because a few participants made a large number of interventions on most trials, this distribution was modestly skewed with a median of 4 and mode of 3. One participant made no interventions at all.

Interventions lasted an average of 3.46 s ($SD = 3.00$) and had a range (the maximum value of the variable during the intervention subtracted from its minimum value) of 138.3 ($SD = 58.89$). This latter measure was strongly bimodal with modes around 100 and 200, indicating that interventions typically consisted of participants dragging a variable from about 0 to one end of the scale (−100 or 100) or then in addition dragging it to the opposite end of the scale. Apart from these large swings, participants typically held the variable steady at a constant value during an intervention. This conclusion is supported by the fact that, within an intervention, the percentage of 100 ms time windows in which the variable had the same value as during the previous window was 71.2%. Four participants had some tendency to "wiggle" the variable through a small range during an intervention but they were the exception.

The interventions were spread about evenly over the three variables. Indeed, all three network variables were manipulated at least once on more than 99% of the trials. Interventions varied modestly as a function of whether the manipulated variable was a cause of other variables in the network. When it was, the intervention was both shorter (3.21 s) and had a narrower range (132.9) than when it wasn't (3.99 s and 149.5), $t_{(28)} = 3.19$ and $t_{(28)} = 6.39$, respectively, both $ps < 0.005$[5]. Apparently, it was easier for participants to identify causes, which involves observing a state change in other network variables, than non-causes, which involves the absence of such changes. Interventions on causes did not vary substantially, in length of time or range of values, as a function of whether they had one or two effects. Interventions also did not vary as a function of whether or not the variable was affected by other variables in the network. In summary, participants recognized that interventions help causal learning, that manipulating all variables is necessary to identify the correct causal structure, and that large interventions are more useful than small ones.

Results Summary

Participants exhibited considerable ability to intervene effectively and learn causal structure in our task. Despite these abilities, they also made systematic errors consistent with the predictions of the LC model. It is not clear whether the data considered as a whole is more consistent with normativity or a more locally focused model. Indeed, it is not even clear that participants are using the OU functional form to infer connections, rather than a more general model, such as one that assumes linearity. For a more granular analysis of people's causal structure learning, we now turn to a number of theoretical accounts of how people learn causal structure.

MODELING

In this task we compare a total of nine models corresponding to different accounts of how people learn causal structure. These accounts can be roughly categorized as modeling people as normative, local, linear, or random in their causal learning behavior. We compare the ability of these models' to predict participants' causal structure judgments.

OU Models

Normative Model

Normative inference for the current task requires that a learner maintain a distributional belief over all possible causal structures and update it according to the data they experience. Equation (6) above defines normative inference in this task. There has been much work suggesting that adults and children are capable learners of causal structures and act roughly in accordance with the normative model, at least in sufficiently simple scenarios (Gopnik et al., 2004; Griffiths and Tenenbaum, 2009). We ask whether these conclusions generalize to the sort of causal systems under investigation here.

Recall that Equation (6) assumes that learners have priors over ω, σ, and the θs. We assume for simplicity that learners acquire a rough approximation of the true values of these parameters [i.e., $\omega = 0.1$, $\sigma = 5$, and $\theta \in (-1, 0, 1)$] while watching the four instructional videos, but assume some spread to accommodate uncertainty. The distributions we assumed over parameters were thus[6]

$$\theta \sim \Gamma(\text{shape} = 5 \times \theta_{true},\ \text{rate} = 5)$$

$$\omega \sim \Gamma(\text{shape} = 100 \times \omega_{true},\ \text{rate} = 100)$$

$$\sigma \sim \Gamma(\text{shape} = 100 \times \sigma_{true},\ \text{rate} = 100)$$

Note that θ values are defined by the graph. For regular connections, θ is distributed as above. For inverse connections, the sampled values are negated. For non-connections θ is 0.

Local Computations Model

We compare the normative model to a "local computations" (LC) model that has been advocated as a general-purpose account of causal learning behavior (Fernbach and Sloman, 2009; Bramley

[5] There were 28° of freedom for these analyses, rather than 29, because one of the 30 participants did not intervene.

[6] The 5 and 95% quantiles associated with these distributions are 0.39 and 1.83 for θ, 0.054 and 0.157 for ω, and 4.64 and 5.37 for σ. These variances were chosen to accommodate a moderate amount of uncertainty in beliefs about each of the higher-level parameters, while still being consistent with the qualitative behavior of the system under the true parameters.

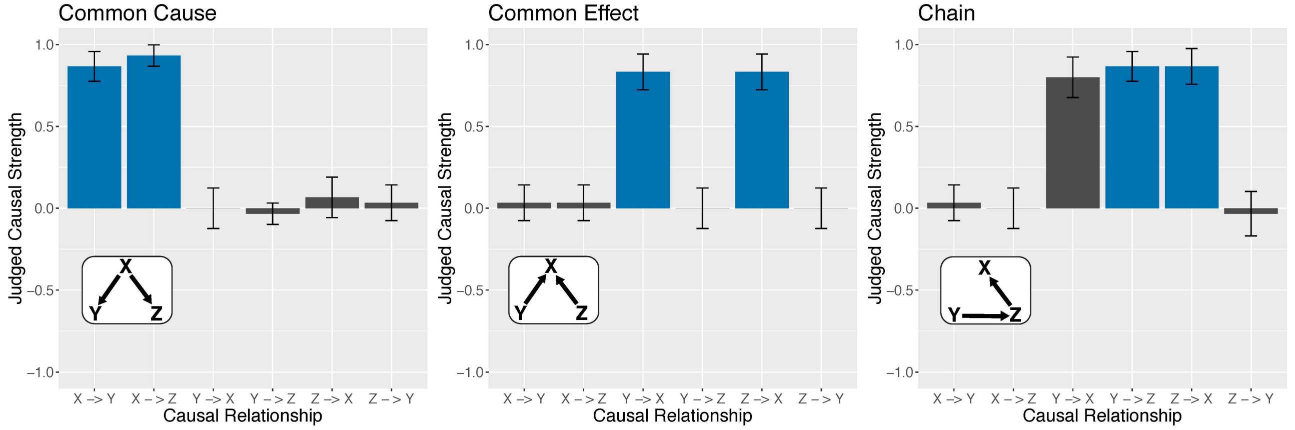


FIGURE 7 | Participant judgments of causal relationships for three tested networks. Bar colors correspond to the true causal structure, namely, blue for regular connections and gray for no connection. Bar heights represent mean θ reported by participants (regular = 1 and none = 0). Because these networks included only regular causal relationships, no instances of inverse relationships are shown. Error bars Denote 95% confidence intervals.

et al., 2017a). Applied to an OU network, the LC model entails deciding, for each potential causal relationship considered in isolation, whether the observed values of those two variables implies a regular, inverted, or zero causal relation. It thus involves applying Equation (5) above to each potential causal relationship. The LC model assumes the same priors over ω, σ, and the θs as the normative model.

A key distinction between the normative and LC models of course is their ability to detect whether a relationship between two variables is mediated by a third. For example, in the network $X \rightarrow Y \rightarrow Z$, X and Z have many of the hallmarks of a direct causal relationship: They are correlated, changes in X precede changes in Z, and intervening on X later affects Z (but not vice versa). Whereas, the normative model would take into account the mediated relationship between X and Z (by noting the absence of an X/Z correlation when controlling for Y), LC, which evaluates individual causal links without consideration of the entire graph, would not recognize the mediating role of Y and so infer $X \rightarrow Z$ in addition to $X \rightarrow Y$ and $Y \rightarrow Z$. Of course, we have already seen partial evidence that participants may be poor at detecting mediated relationships (**Figure 7**). Modeling will reveal whether the LC model is a good account of all the data, or if it only accounts for participants' errors.

Alternative Models

We compare the two OU-based models to alternatives that assume linear relationships between cause and effect. In particular, we compare two approaches to modeling timeseries information from the literature: time-lagged correlation and Granger causality. Each of these approaches is applied to three candidate representations for learning causal structure between continuous variables, as introduced by Soo and Rottman (2018); *state representations*, *difference scores*, and *trinarized difference scores*.

In these linear models, the value of variable i at time t is modeled as

$$P(v_i^t | v^{t-1}, \sigma, \beta) = \sum_j \left[\beta_{ji} \cdot v_j^{t-1} \right] + N(0, \sigma) \tag{7}$$

where j denotes all causes of variable i (including i itself) and β_{ji} denotes the partial slope coefficient or strength of that cause on the effect. Analogously to our treatment of θ values in the OU models, for the linear models we assume some uncertainty about the strength parameter $p(\beta)$ but that these differ in sign for regular and inverse connections, and also model people as having uncertainty over standard deviation $p(\sigma)$. The marginal likelihood of v_i for a graph thus involves computing, for each timepoint, the likelihood of that variable's value given the β predictors defined by the graph and the value(s) of its cause(s), and marginalizing over $p(\beta)$ and $p(\sigma)$. We treat interventions in the same manner as the OU models. As before, we compute the total likelihood as the product of the marginal likelihoods of all variables at all timepoints under each graph, assume an initially uniform prior over graphs and compute the resulting posterior. The unnormalized posterior probability of a causal graph given all values of all variables at all timepoints is thus

$$P(g_k | v; \iota) \propto \prod_t \prod_i \int_\beta \int_\sigma P(v_i^t | v^{t-1}, \sigma, \beta; \iota_i) P(\beta | g_k) P(g_k) P(\sigma) \, d\sigma \, d\beta \tag{8}$$

This general procedure can be applied to each of the linear models by modifying the state representation v or prior over β. For the three candidate representations introduced by Soo and Rottman (2018): State representations involves inference over the actual variable values; difference scores involves inference over variable values after computing $v^t - v^{t-1}$; trinarized difference scores involves inference over difference scores that have been converted to -1 when negative and 1 when positive.

The difference between time-lagged correlation and Granger causality is just whether β_{ii} is included as a predictor, that is, whether v_i^t is influenced by v_i^{t-1} as well as its causes. Granger causality includes this term while Time-lagged correlation does not.

Unlike the OU models, there is no natural ground truth parametrization for the linear models on which to center reasonable distributional parameter beliefs. Thus, we must find another way to choose reasonable settings for $p(\beta)$ and $p(\sigma)$. We chose the mean of our distributions by fitting the $\hat{\beta}_{ii}$, $\hat{\beta}_{ji}$, and $\hat{\sigma}$ values that maximized the posterior probability of the true causal graphs across all subject data (including β_{ii} for the Granger models). We then made analogous assumptions about the spread around these means as we did for θ and σ in the OU models—namely,

$$\beta \sim \Gamma(\text{shape} = 5 \times \hat{\beta},\ \text{rate} = 5)$$
$$\sigma \sim \Gamma(\text{shape} = 100 \times \hat{\sigma},\ \text{rate} = 100).$$

β values are treated the same as in the OU models. Regular connections are distributed as above, inverse connections are negated.

Comparing the Models

We compare participants' structure judgments to the predictions of these models across all the test trials in our experiment. In total, we consider nine models. These are eight described above: (1) *normative*, (2) *local computations (LC)*, and three variants of both (3–5) *Granger causality* and (6–8) *Time lagged correlation* varying whether they were based directly on states, difference scores, or trinarized difference scores. Finally, we compare these against (9) a *Baseline* model that assumes each judgment is a random selection from the space of possible graphs. We marginalized over θ, ω, σ by drawing 1,000 samples from their respective distributions and averaging the likelihood within each causal model. To account for decision noise in selecting causal graphs from their posterior distributions, for each model apart from the baseline we fit (by maximum likelihood using R's optim function) a single softmax parameter τ that maximized the posterior probability of participant selections.

Results and Discussion

Table 1 details the results of our comparison. For each inference model we report the overall proportion of the true connections identified across all trials assuming the most probable graph is selected at the end of each trial (Accuracy column), the proportion of participant's edge judgments that correspond with the most probable graph under the model (Judge column), the Bayesian Information Criterion of all participant's judgments according to that model (BIC column); and the number of participants best fit by each model[7].

Unsurprisingly, the normative model was the most successful at recovering the underlying structure, but many other models were also successful. The only models that struggled were those that used trinarized difference scores as their representation, showing that the magnitude of changes in the variables is important to capturing the structure of the data.

Next, we compared the maximum *a posteriori* estimates of causal structure of the models to participant judgments. In this coarse measure, the OU models were roughly equal to each other in matching participant judgments, and were also similar to some of the linear models.

The results of the more sensitive posterior probability analysis were clearer in distinguishing between models. Over all participants, the LC model had the highest log-likelihood. On a per participant basis, of the 30 participants 21 were best fit by the LC model, with the normative model being the best account of four participants. The remaining five participants were split among the linear models or were at baseline.

GENERAL DISCUSSION

In this paper, we introduced a generative model of causal influence relating continuous variables over time. We showed how such systems can exhibit emergent behaviors, such as excitatory or inhibitory feedback and oscillations, depending on specific settings of relative causal strengths between variables. When learning from this rich data, people were best described as considering individual pairs of variables, rather than updating their beliefs over entire structures. This finding accords with an intuitive description of how people handle continuous information flowing in real time: they focus their attention on smaller, more manageable problems rather than attempting to tackle the full torrent of information.

Local Inference

A key result in our task was that most participants evaluated pairwise relationships between variables rather than updating their beliefs over all possible causal structures. This conclusion was drawn from the superior fit of the locally focused LC model, and corroborated by qualitative results, such as the finding that participants often inferred direct causal relationships between variables that were in fact only indirectly related (through a third mediating variable). These results are consistent with previous findings suggesting that, rather than representing a full hypothesis space, people tend to consider a single hypothesis to which they make small alterations (Quine, 1960; Fernbach and Sloman, 2009; Bramley et al., 2017a). Here we show that this principle of causal learning extends to much richer scenarios. Indeed, it may be the case that real time continuous information places stronger demands on attention and memory than the original settings that provided evidence for the LC model. If this were true, it would be especially reasonable to use the resource-efficient local strategy in these more demanding environments.

A potential alternative conceptualization of the LC model is that it instantiates the idea that distal causes are still considered as causal. For example, most people would not find it inappropriate to say that the reintroduction of wolves to Yellowstone National Park caused changes to the ecosystem, even if many of these changes came indirectly through other variables, such as changes

[7] A *post-hoc* power test was computed for the null hypothesis that the number of participants best fit by each of the nine models would be equally distributed. For a chi-squared test with 30 participants, 8° of freedom, and the observed effect size of 1.91, the probability of observing an $\alpha < 0.05$ is 0.999.

TABLE 1 | Summary of model accuracy and performance.

	Model	State representation	Accuracy	Judge	BIC	Px
1	OU local computations		0.89	0.82	6,163	21
2	OU normative		1.00	0.82	6,475	4
3	Granger causality	States	0.91	0.78	7,079	1
4		Difference scores	0.82	0.69	8,415	1
5		Trinarized diff scores	0.49	0.42	9,859	0
6	Time-lagged correlation	States	0.89	0.74	7,901	1
7		Difference scores	0.82	0.69	8,407	0
8		Trinarized diff scores	0.63	0.50	9,793	0
9	Baseline		0.17	0.17	9,888	2

Accuracy, *proportion of links drawn that match ground truth;* Judge, *proportion of links drawn that match participant judgments;* BIC, *Bayesian Information Criterion;* Px, *number of participants best fit by that model.*

in the movement of elk (Fortin et al., 2005). While this is a reasonable conceptualization, we believe that it is not as good an account of our data as the LC model. For one, we explicitly provided participants with an example in the instructions that showed the movement of a chain network without the additional indirect connection. This should have reduced the possibility that participants were unclear about whether they should consider distal causes as causal. This accords with findings in the literature that people exhibit locality despite feedback, incentives, and explicit instruction with examples that encourage people to not draw the additional causal link (Fernbach and Sloman, 2009; Bramley et al., 2015, 2017a). More fundamentally, this "distal" account makes assumptions about how people are approaching the task that we consider unlikely. It models them as doing full normative inference, and then having a response bias to draw indirect connections. **Figure S1** shows that indirect connections were less likely to be responded to as causal than the direct connections, which would imply a response bias where participants have the full causal model but would only on occasion draw the additional indirect connection. The LC model, in contrast, naturally considers indirect connections as less causal due to the underlying dynamics of OU networks. While indirect causal relationships do have many hallmarks of direct causal relationships (correlation, temporal asymmetry, asymmetric results of interventions), they are not identical. In $X \rightarrow Y \rightarrow Z$, changes to Z in response to X are more temporally removed and noisier than would be predicted if there were a direct $X \rightarrow Z$ connection, and therefore the LC model assigns a lower (but still reliably non-zero) probability to these potential connections. Because the LC model accounts for the patterns of errors as naturally arising from the interaction of system dynamics and cognitive limitations, rather than as a response bias over normative inference, we consider it a better account of the behavior of participants in our task.

Interventions

One contribution of the OU network framework is the introduction of a qualitatively different type of intervention. In a typical study of causal cognition learners are able to, on a particular trial, turn a variable on or off and observe the values of other variables. In contrast, interventions in our task are extended through time and can encompass a wide range of variable values. Participants generally recognized that the most informative actions involved large swings in variable values and systematic manipulation of each variable in the system[8].

Nevertheless, note that while their interventions were informative they were less than optimal. In fact, the most efficient interventions in this task involve rapid swings between the ends of the variable's range. But whereas participants used the full range, they tended to hold a variable at one value for longer than necessary. Doing so yields useful but somewhat redundant information. Of course, perhaps this strategy reflected participants' need for redundant information imposed by cognitive processing limits. It may also reflect their inability or unwillingness to engage in the rapid motor movements required by the optimal strategy.

Although participants could intervene on any variable at any time to set it to any value, they were constrained to manipulating one variable at a time. Future studies could expand the action space by, for example, allowing participants to "freeze" one variable at a value while manipulating others. Of course, an ability to "control for" one variable while investigating the relationship between two others might help learns identify mediating relationships. For example, freezing Y and then manipulating X in $X \rightarrow Y \rightarrow Z$ would result in to no change in Z, perhaps reducing the chance that the learner would conclude $X \rightarrow Z$. This approach could be considered an application of learning strategies from the CPS literature to environments without sharp distinctions between input and output nodes (Kuhn and Brannock, 1977; Schoppek and Fischer, 2017), with the additional information generated by the "Do()" operator's graph surgery.

Future Directions

The proposed OU network framework can be extended across a variety of dimensions in future research. For example, in this paper's instantiation of OU networks, a cause impacts an effect on the next timepoint. The impact of a cause on effect could be distributed over multiple timepoints, or at some stochastically selected timepoint. Such studies could contribute to debates

[8] The observed systematic strategy of manipulating a single variable, holding it at a value, and observing the downstream effects closely corresponds to successful learning strategies from the CPS literature, such as VOTAT and PULSE.

about the influence of time on causal learning, such as that judgments of causality are strengthened by temporal contiguity (Shanks et al., 1989) or the reliability of delays (Buehner and May, 2003; Bramley et al., 2018). Varying the gap between timepoints (in this task t to t+1 was 100ms) may result in different approaches by participants. Use of continuous variables naturally allows consideration of a greater number functional forms relating causes and effects (Griffiths and Tenenbaum, 2009). Latent causes can be introduced to model implicit inference of mechanisms relating cause and effect. Complex, non-linear data can be generated to study people's learning from time series data (Soo and Rottman, 2018; Caddick and Rottman, 2019). The outcomes of experiments using these richer causal systems will help to evaluate the generalizability of models of causal cognition that have heretofore been tested mostly on Bayes nets applied to discrete events.

The formalism developed in this paper also has potential application to the domain of control. Many aspects of everyday life, as well as interesting domains in AI and machine learning, can be can be classed as control problems in which there is initial or ongoing uncertainty about the structure of the control domain. As discussed in the introduction, there is an extensive literature known as Complex Problem Solving that has participants manipulate environments that are reactive to their decisions to maximize gain (for review, see Osman, 2010). One limitation of extant work is that they do not include learning models that can help distinguish between learning and control performance. In parallel, much recent attention in machine learning has been given to demonstrations of successful control in small worlds, such as atari and board games. However, generalization to new goals or related environments continues to be poor (Lake et al., 2017). In recent work, we propose OU networks as a systematic class of control environments. This approach allows research into human control to ask new questions, such as what structures are inherently easy or hard to identify or control and under what circumstances does successful control depend on an accurate model of a system's structure (Davis et al., 2018).

Functional Form

Given people's well-known bias toward assuming linear functional forms (Brehmer, 1974; Byun, 1996; DeLosh et al., 1997; Kalish et al., 2004, 2007; Kwantes and Neal, 2006), it may be a surprising result that the alternative models assuming linearity did not match people's judgments as well as those using the Ornstein–Uhlenbeck functional form. This result has a number of possible explanations. For one, as discussed before, Ornstein–Uhlenbeck processes appear to be relatively common across a range of domains, and people may have a developed representation of the functional form that they brought to the task. It is also possible that participants do not have a direct representation of Ornstein–Uhlenbeck processes, but were able to recognize higher-order movement statistics that are not present in linear models (e.g., OU processes, unlike linear relationships, exhibit acceleration toward their attractor basin). For example, people may have applied a general function approximator, such as a Gaussian Process to the relationship between cause and effect and abstracted a function closer to OU processes than linearity. Future work could explore settings where learning the functional form between cause and effect is not possible (such as one-shot learning) or settings where the impact a cause has on its effect is linear.

Limitations

There are a number of limitations to the current project that could be addressed with further experiments. For one, while we did account for uncertainty over parameters of our models, we did not account for other sources of noise, such as the likelihood that people cannot attend to all three variables simultaneously[9]. This issue will likely compound as more variables are added. Additionally, the presented analyses in this paper discuss but do not model intervention decision-making, a critical component of the active learning of causal structure. Future analyses would naturally involve, as a benchmark to compare against humans, models for selecting actions that maximize expected information gain. This information maximizing strategy could be compared to other strategies from the Complex Problem Solving literature that involve changing a single variable at a time (Kuhn and Brannock, 1977; Schoppek and Fischer, 2017).

Conclusions

We have no doubt that the canonical causal relationships between discrete events (e.g., take a pill → headache relieved) that have been the main focus of causal cognition often serve as highly useful and approximately correct parts of human's semantic representation of the world. But sometimes details matter. Causal influences emerge over time, may reflect functional relationships that are as complex as the underlying mechanisms that produce them, and afford interventions that vary in their duration and intensity. Complex patterns of feedback may be the rule rather than the exception (Cartwright, 2004; Strevens, 2013; Sloman and Lagnado, 2015). Apprehending these properties may even be a precondition to forming the (highly summarized and approximate) causal relations between discrete events that are so simple to represent and easy to communicate.

We instantiated a learning task in which people were confronted with some of these challenges, including continuously-observed continuous variables, feedback cycles, and the ability to carry out extended interventions. We found that they exhibited considerable success identifying the correct causal structure but also committed systematic errors, errors consistent with a model that describes people as narrowly investigating individual causal relationships rather than updating their beliefs wholesale. We hope that the formalism presented in this paper will be help spur greater study of the mechanisms for learning and action in this important class of problems.

[9] Although Vul et al.'s (2009) finding that people optimally allocate attention to particles moving according to an OU process may ameliorate the latter concern.

DATA AVAILABILITY STATEMENT

Raw data is available at the public website: https://zach-davis.github.io/publication/cvct/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by New York Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

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

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00244/full#supplementary-material

REFERENCES

Ali, N., Chater, N., and Oaksford, M. (2011). The mental representation of causal conditional reasoning: mental models or causal models. *Cognition* 119, 403–418. doi: 10.1016/j.cognition.2011.02.005

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. New York, NY: Cambridge University Press.

Berry, D. C., and Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Q. J. Exp. Psychol. Sect. A* 36, 209–231.

Bramley, N. R., Dayan, P., Griffiths, T. L., and Lagnado, D. A. (2017a). Formalizing neurath's ship: approximate algorithms for online causal learning. *Psychol. Rev.* 124:301. doi: 10.1037/rev0000061

Bramley, N. R., Gerstenberg, T., and Lagnado, D. (2014). "The order of things: inferring causal structure from temporal patterns," in *Proceedings of the Annual Meeting of the Cognitive Science Society* (Quebec City, QC) Vol. 36.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., and Lagnado, D. A. (2018). Time in causal structure learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 44:1880. doi: 10.1037/xlm0000548

Bramley, N. R., Lagnado, D. A., and Speekenbrink, M. (2015). Conservative forgetful scholars: how people learn causal structure through interventions. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 708–731. doi: 10.1037/xlm0000061

Bramley, N. R., Mayrhofer, R., Gerstenberg, T., and Lagnado, D. A. (2017b). "Causal learning from interventions and dynamics in continuous time," in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (Austin, TX: Cognitive Science Society).

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organ. Behav. Hum. Perform.* 11, 1–27.

Brehmer, B., and Allard, R. (1991). "Dynamic decision making: the effects of task complexity and feedback delay," in *New Technologies and Work. Distributed Decision Making: Cognitive Models for Cooperative Work*, eds J. Rasmussen, B. Brehmer, and J. Leplat (Chichester: John Wiley & Sons), 319–334.

Buehner, M. J., and May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *Q. J. Exp. Psychol. Sect. A* 56, 865–890. doi: 10.1080/02724980244000675

Burns, P., and McCormack, T. (2009). Temporal information and children's and adults' causal inferences. *Think. Reason.* 15, 167–196. doi: 10.1080/13546780902743609

Byun, E. (1996). *Interaction between prior knowledge and type of nonlinear relationship on function learning* (Ph.D. thesis), Lafayette, IN: ProQuest Information & Learning.

Caddick, Z. A., and Rottman, B. M. (2019). "Politically motivated causal evaluations of economic performance," in *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (Montreal, CA: Cognitive Science Society).

Cartwright, N. (2004). Causation: one word, many things. *Philos. Sci.* 71, 805–819. doi: 10.1086/426771

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychol. Rev.* 104:367.

Coenen, A., Rehder, B., and Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cogn. Psychol.* 79, 102–133. doi: 10.1016/j.cogpsych.2015.02.004

Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE* 8:e57410. doi: 10.1371/journal.pone.0057410

Davis, Z., Bramley, N., Rehder, B., and Gureckis, T. M. (2018). "A causal model approach to dynamic control," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (Madison, WI: Cognitive Science Society).

DeLosh, E. L., Busemeyer, J. R., and McDaniel, M. A. (1997). Extrapolation: the sine qua non for abstraction in function learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 23:968.

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

Fernbach, P. M., and Sloman, S. A. (2009). Causal learning with local computations. *J. Exp. Psychol. Learn. Mem. Cogn.* 35:678. doi: 10.1037/a0014928

Fortin, D., Beyer, H. L., Boyce, M. S., Smith, D. W., Duchesne, T., and Mao, J. S. (2005). Wolves influence elk movements: behavior shapes a trophic cascade in Yellowstone National Park. *Ecology* 86, 1320–1330. doi: 10.1890/04-0953

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason.* 7, 69–89. doi: 10.1080/13546780042000046

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychol. Rev.* 111:3. doi: 10.1037/0033-295X.111.1.3

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., and Martin, R. (2013). A multitrait–multimethod study of assessment instruments for complex problem solving. *Intelligence* 41, 579–596. doi: 10.1016/j.intell.2013.07.012

Greiff, S., Niepel, C., Scherer, R., and Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files. *Comput. Hum. Behav.* 61, 36–46. doi: 10.1016/j.chb.2016.02.095

Greville, W. J., and Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *J. Exp. Psychol. Gen.* 139:756. doi: 10.1037/a0020976

Griffiths, T. L. (2004). *Causes, coincidences, and theories* (Ph.D. thesis), Stanford University, Stanford, CA, United States.

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

Griffiths, T. L., and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychol. Rev.* 116:661. doi: 10.1037/a0017201

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., et al. (2016). Psiturk: an open-source framework for conducting replicable behavioral experiments online. *Behav. Res. Methods* 48, 829–842. doi: 10.3758/s13428-015-0642-8

Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., and Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition* 133, 611–620. doi: 10.1016/j.cognition.2014.08.011

Hitchcock, C. (2018). "Probabilistic causation," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Metaphysics Research Lab, Stanford

University). Available online at: https://plato.stanford.edu/archives/fall2018/entries/causation-probabilistic/

Hume, D. (1959). *Enquiry Concerning Human Understanding*. New York, NY: Dover (Original Work Published 1748).

Kalish, M. L., Griffiths, T. L., and Lewandowsky, S. (2007). Iterated learning: intergenerational knowledge transmission reveals inductive biases. *Psychon. Bull. Rev.* 14, 288–294. doi: 10.3758/BF03194066

Kalish, M. L., Lewandowsky, S., and Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychol. Rev.* 111:1072. doi: 10.1037/0033-295X.111.4.1072

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen.* 136:430. doi: 10.1037/0096-3445.136.3.430

Kuhn, D., and Brannock, J. (1977). Development of the isolation of variables scheme in experimental and "natural experiment" contexts. *Dev. Psychol.* 13:9.

Kwantes, P. J., and Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *J. Exp. Psychol. Learn. Mem. Cogn.* 32:1019. doi: 10.1037/0278-7393.32.5.1019

Lacko, V. (2012). Planning of experiments for a nonautonomous Ornstein-Uhlenbeck process. *Tatra Mount. Math. Publ.* 51, 101–113. doi: 10.2478/v10127-012-0011-2

Lagnado, D. A., and Sloman, S. A. (2006). Time as a guide to cause. *J. Exp. Psychol. Learn. Mem. Cogn.* 32:451. doi: 10.1037/0278-7393.32.3.451

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40, 1–25. doi: 10.1017/S0140525X16001837

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., and Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychol. Rev.* 115:955. doi: 10.1037/a0013256

Marsh, J. K., and Ahn, W.-k. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *J. Exp. Psychol. Learn. Mem. Cogn.* 35:334. doi: 10.1037/a0014929

McCormack, T., Frosch, C., Patrick, F., and Lagnado, D. (2015). Temporal and statistical information in causal structure learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 41:395. doi: 10.1037/a0038385

Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychol. Bull.* 136:65. doi: 10.1037/a0017815

Pacer, M. D., and Griffiths, T. L. (2011). "A rational model of causal induction with continuous causes," in *Proceedings of the 24th International Conference on Neural Information Processing Systems* (Granada: Curran Associates Inc.), 2384–2392.

Pacer, M. D., and Griffiths, T. L. (2012). "Elements of a rational framework for continuous-time causal induction," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34 (Sapporo).

Pearl, J. (2009). *Causality*. New York, NY: Cambridge University Press.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cogn. Psychol.* 72, 54–107. doi: 10.1016/j.cogpsych.2014.02.002

Rothe, A., Deverett, B., Mayrhofer, R., and Kemp, C. (2018). Successful structure learning from observational data. *Cognition* 179, 266–297. doi: 10.1016/j.cognition.2018.06.003

Rottman, B. M., and Keil, F. C. (2012). Causal structure learning over time: observations and interventions. *Cogn. Psychol.* 64, 93–125. doi: 10.1016/j.cogpsych.2011.10.003

Schoppek, W., and Fischer, A. (2017). Common process demands of two complex dynamic control tasks: transfer is mediated by comprehensive strategies. *Front. Psychol.* 8:2145. doi: 10.3389/fpsyg.2017.02145

Shanks, D. R., Pearson, S. M., and Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *Q. J. Exp. Psychol.* 41, 139–159.

Sloman, S. A. (2005). *Causal Models: How People Think About the World and Its Alternatives*. New York, NY: Oxford University Press.

Sloman, S. A., and Lagnado, D. (2005). Do we "do". *Cogn. Sci.* 29, 5–39. doi: 10.1207/s15516709cog2901_2

Sloman, S. A., and Lagnado, D. (2015). Causality in thought. *Annu. Rev. Psychol.* 66, 223–247. doi: 10.1146/annurev-psych-010814-015135

Soo, K. W., and Rottman, B. M. (2018). Causal strength induction from time series data. *J. Exp. Psychol. Gen.* 147:485. doi: 10.1037/xge0000423

Stadler, M., Fischer, F., and Greiff, S. (2019). Taking a closer look: an exploratory analysis of successful and unsuccessful strategy use in complex problems. *Front. Psychol.* 10:777. doi: 10.3389/fpsyg.2019.00777

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., and Blum, B. (2003). Inferring causal networks from observations and interventions. *Cogn. Sci.* 27, 453–489. doi: 10.1207/s15516709cog2703_6

Strevens, M. (2013). Causality reunified. *Erkenntnis* 78, 299–320. doi: 10.1007/s10670-013-9514-8

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285.

Taylor, E. G., and Ahn, W.-k. (2012). Causal imprinting in causal structure learning. *Cogn. Psychol.* 65, 381–413. doi: 10.1016/j.cogpsych.2012.07.001

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10.

Uhlenbeck, G. E., and Ornstein, L. S. (1930). On the theory of the brownian motion. *Phys. Rev.* 36:823.

Vollmeyer, R., Burns, B. D., and Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cogn. Sci.* 20, 75–100.

Vul, E., Alvarez, G., Tenenbaum, J. B., and Black, M. J. (2009). "Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model," in *Advances in Neural Information Processing Systems 22*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Curran Associates, Inc.), 1955–1963.

Waldmann, M. R., and Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.* 31:216. doi: 10.1037/0278-7393.31.2.216

Waldmann, M. R., and Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *J. Exp. Psychol. Gen.* 121, 222–236.

Conflict of Interest: 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.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 07 April 2020
doi: 10.3389/fpsyg.2020.00630

Events and Causal Mappings Modeled in Conceptual Spaces

Peter Gärdenfors[1,2]*

[1] Department of Philosophy and Cognitive Science, Lund University, Lund, Sweden, [2] Palaeo-Research Institute, Faculty of Humanities, University of Johannesburg, Johannesburg, South Africa

The aim of the article is to present a model of causal relations that is based on what is known about human causal reasoning and that forms guidelines for implementations in robots. I argue for two theses concerning human cognition. The first is that human causal cognition, in contrast to that of other animals, is based on the understanding of the *forces* that are involved. The second thesis is that humans think about causality in terms of *events*. I present a two-vector model of events, developed by Gärdenfors and Warglien, which states that an event is represented in terms of two main components – the force of an *action* that drives the event, and the *result* of its application. Apart from the causal mapping, the event model contains representations of a patient, an agent, and possibly some other roles. Agents and patients are objects (animate or inanimate) that have different properties. Following my theory of conceptual spaces, they can be described as vectors of property values. At least two spaces are needed to describe an event, an action space and a result space. The result of an event is modeled as a vector representing the change of properties of the patient before and after the event. In robotics the focus has been on describing results. The proposed model also includes the causal part of events, typically described as an action. A central part of an event category is the mapping from actions to results. This mapping contains the central information about *causal* relations. In applications of the two-vector model, the central problem is how the event mapping can be learned in a way that is amenable to implementations in robots. Three processes are central for event cognition: causal thinking, control of action and learning by generalization. Although it is not yet clear which is the best way to model how the mappings can be learned, they should be constrained by three corresponding mathematical properties: monotonicity (related to qualitative causal thinking); continuity (plays a key role in activities of action control); and convexity (facilitates generalization and the categorization of events). I argue that Bayesian models are not suitable for these purposes, but some more geometrically oriented approach to event mappings should be used.

Keywords: causation, robotics, events, action, conceptual space

Edited by:
York Hagmayer,
University of Göttingen, Germany

Reviewed by:
Igor Douven,
Université Paris-Sorbonne, France
Ralf Mayrhofer,
University of Göttingen, Germany

***Correspondence:**
Peter Gärdenfors
Peter.Gardenfors@lucs.lu.se;
pg22350@gmail.com

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

Received: *22 September 2019*
Accepted: *16 March 2020*
Published: *07 April 2020*

Citation:
Gärdenfors P (2020) Events and Causal Mappings Modeled in Conceptual Spaces. Front. Psychol. 11:630. doi: 10.3389/fpsyg.2020.00630

INTRODUCTION

Causal reasoning is a central cognitive competency, allowing us to reliably, albeit not perfectly, predict the future and to understand the causes of events that we observe. This form of reasoning has been studied extensively in psychology and philosophy (see e.g., Waldmann and Hagmayer (2013) for an overview). In this article, my focus will be on aspects of human causal reasoning that should be considered when developing robotic systems that are capable of similar forms of reasoning.

If we want to develop efficient systems for human-robot interaction, the best way is to have robots reason about causes in the same way as humans do. Therefore, we need a model of human causal cognition that allows implementation. Pearl (2018) writes that we recognize human reasoning "through words such as 'preventing,' 'cause,' 'attributed to,' 'discrimination,' and 'should I.' Such words are common in everyday language, and our society constantly demands answers to such questions. Yet, until very recently science gave us no means even to articulate them, let alone answer them. Unlike the rules of geometry, mechanics, optics or probabilities, the rules of cause and effect have been denied the benefits of mathematical analysis."

This article will argue for two theses concerning human cognition. The first is that causal cognition is based on the understanding of the *forces* that are involved. In the section Causal Reasoning with Forces, I present some data concerning the differences between human causal reasoning and that of other animals. I propose that the best way to understand these differences is that humans have evolved mental representations of the forces behind an action or a physical process that lead to an effect.

The second thesis is that humans think about causality in terms of *events*[1]. However, unlike other models in philosophy and psychology where causality is seen as a relation between events, the model presented here moves causality inside events in the sense that an event is modeled as containing two vectors representing a cause as well as a result. In Section 3, I present a model that is based on a mapping from actions to results. The purpose of such a mapping is to represent causal relations. Actions are modeled in terms of forces, while effects are modelled as different kinds of changes, for example, a change in the physical location or a change of some property of the agent. Apart from the causal mapping, the event model contains representations of an agent, a patient and possibly some other roles.

Three cognitive processes crucially depend on event cognition: causal thinking, control of action and learning by generalization. All three processes are important for robot applications. The central problem is to model the event mapping and how it is learned in a way that is amenable to implementation.

The mapping from forces to results may have a complicated structure due to context dependent or unknown counterforces. However, the mapping is constrained by three properties that correspond to the three cognitive processes respectively: (1) Larger forces lead to larger results (related to qualitative causal thinking); (2) small changes in the force lead to small changes in the result (plays a key role in action control); and (3) intermediate results are caused by intermediate forces (facilitates generalization and the categorization of events). These properties will be presented and analyzed in the section Three Constraints on the Causal Mapping.

On the basis of the event model and the constraints on the causal mapping, I will discuss some ideas about how such mappings can be handled in a robot. This will be the topic of in the section Implementing the Event Model and the Causal Mapping in Robots. The main problem to be solved is how the event mapping from causes to effects can be learned. Here the three constraints turn out to be central. I also argue that Bayesian models are not appropriate since they cannot account for the three constraints on the causal mapping in a natural way.

CAUSAL REASONING WITH FORCES

Human Reasoning About Forces

The sensory influx to the human brain is extremely rich – a "blooming buzzing confusion" according to James (1890, p. 42). It is something of a wonder that the brain can sort up the information received by our senses. In particular, it has a capacity to discover causal relations between complex phenomena. It is, however, still largely an open question how this mechanism works.

There are several proposals for how to analyze causal cognition. Gärdenfors (2003, Section 2.8) distinguishes between four kinds of causal reasoning: (a) Being able to foresee the physical effects of one's own actions (the first type to develop in infants); (b) being able to foresee the effects of others' actions; (c) understanding the causes of others' actions; and (d) understanding the causes of physical events. Along similar lines, Woodward (2011) distinguishes between *egocentric learning*, which is the ability to learn that one's own physical actions can cause certain outcomes. The second kind is *agent causal learning*, when one also learns about cause from the actions of others. The third kind is *observation/action causal learning*, when one is able to integrate a natural signs or patterns with the other two types of learning[2].

The models indicate that being able to categorize actions is a necessary prerequisite for understanding causal relations. Psychological studies have established that the brain processes lead to a considerable information reduction when actions are classified. For example, Johansson (1973) showed that the kinematics of a movement contain is sufficient to categorize an action. He attached light bulbs to the joints of actors who were dressed in black and moved against a black background. The actors were then filmed while performing bodily actions such as walking, running and dancing. When subjects saw the movies, in which only the dots of light could be perceived, they correctly categorized the actions within a few hundred milliseconds.

The upshot of these experiments is that the kinematics of a movement contains information that is sufficient for the identification of the underlying dynamic force patterns, that is velocities and accelerations (Runesson, 1994). Further psychological evidence [Wolff (2007, 2008), Wolff and Shepard (2013), Wolff and Thorstad (2017)] supports that people can directly perceive the *forces* that control different kinds of motion.

[1] Davidson (1967, p. 179) writes that "events have a unique position in the framework of causal relations."

[2] A more detailed classification is presented by Lombard and Gärdenfors (2017) and Gärdenfors et al. (2018).

In other words, the sensory input generated by the movements of an individual (or an object) is sufficient for the brain to calculate the forces that lead to the movements. The process is automatic: people cannot help but seeing the forces.

In the philosophical literature, a cause has mainly been viewed as something that makes a difference with respect to some effect. The differences are typically analyzed in terms of co-variations (see Waldmann and Hagmayer, 2013 for a presentation). However, nothing is said about *how* it makes a difference. Theories of causation that are based on forces provide an explanation (Wolff, 2007). Forces also open up for new empirical methods to study casual relations that go beyond covariations.

The capacity to understand the role of physical forces, not just forces involved in animal actions, develops early in human infants. Michotte (1963) showed that if one object moving on a screen collided with another object and the other object started moving in the same direction, then adults perceived the launching of the second object as caused by the movement of the first. In contrast, if the second object only started moving half a second after the collision, then the delay destroyed the impression of causality. Leslie and Keeble (1987) performed Michotte's experiments with six-month-old infants and showed that they reacted differently to the two types of events. Leslie (1995) concludes that infants have a special system in their brains for mapping the 'forces' of objects.

Animal Reasoning About Forces

It seems that non-human primate reasoning about forces is less developed compared to that of humans. For example, in his early experiments on chimpanzee planning, Köhler (1917) observed that apes had great difficulties in stacking boxes on top of each other. He notes about Sultan, the best problem solver among the chimpanzees, that when he tried to put a second box on top of a first, "instead of placing it on top of the first, as might seem obvious, began to gesticulate with it, . . . he put it beside the first, then in the air diagonally above, and so forth." After similar observations on other apes, Köhler (1917, p. 149) concludes that "there is practically no statics to be noted in the chimpanzee." For more experiments in the same direction see Tomonaga et al. (2007) and Cacchione et al. (2009). These observations indicate that apes in general do not have a well-developed understanding of the role of gravitation on other objects than their own bodies.

Povinelli (2000) also performed a series of experiments indicating that chimpanzees and other primates are very limited in their capacities to reason about gravitation. These experiments have been followed by a series of others (e.g., Call, 2010; Hanus and Call, 2008; Martin-Ordas et al., 2008; Penn and Povinelli, 2007), and they have generated an extended debate (see Seed and Call, 2009; Seed et al., 2011). Povinelli and Penn (2011, p. 77) conclude that "only humans are capable of second-order relational reasoning, and only humans, therefore, have the cognitive machinery that can support higher-order, theory-like, causal relations." In line with this, Johnson-Frey (2003: 201) writes: "Comparative studies of chimpanzee tool use indicate that critical differences are likely to be found in mechanisms involved in causal reasoning rather than those implementing sensorimotor transformations."

Furthermore, in a comparative study of on nut-cracking in humans and chimpanzees (Boesch et al., 2017), it was found that humans understood how to apply force to extract numerous nut species through using hammerstones. Yet, the chimpanzees only ever applied such force to Panda nuts, even though they regularly eat hard Irvingia nuts using their teeth. This is a good example of how humans, compared to chimpanzees, have a more abstract causal understanding of tool-assisted force application, allowing us to apply similar solutions to a wider range of subsistence problems. By adding the ability to mentally represent detached forces – and not just actions – as causes, the human mind evolved to extend its capacities to reason and to plan beyond that of other primate species. Gärdenfors and Lombard (submitted) argue that this development was driven (at least in part) by more advanced tool use and manufacturing.

A Cognitive Approach to Causation

This comparison between the causal reasoning of humans and other animals provides a reason for focusing on models that are based on forces also in developing causal reasoning in robots. In the following section I present a model that can function as a framework for computational implementations.

The basic ontological position of my approach to causal reasoning is that causes are cognitive constructions and not relations in the real world. In other words, my account is cognitivist rather than realist. For an argument for this position see Wolff (2007, p 7).

Another central aspect is that the forces of an agent are not the only elements involved in human causal judgements, but counterforces of various kinds (forces exerted by a patient or contextual forces such as gravitation) are also taken into account. This aspect is included in Talmy's (1988) 'force dynamics' and is further developed in Wolff's (2007, 2008, 2012) 'dynamics model'. Wolff (2007) has shown that adults can combine different kinds of forces in their reasoning. For example, they can estimate the combined forces of a boat motor and the wind and their effects on how the boat crosses a lake. Depending on how the 'affector' force vector (produced by an agent) combines with a 'patient' force vector to generate a 'result' vector, subjects judge that the affector force either *causes*, *enables* or *prevents* an effect. These results indicate that subjects cognitively distinguish between different kinds of causal relations. Talmy's force dynamics is grounded in physical events, but it is also used to understand psychological or social interactions.

Göksun et al. (2013) extended Wolff's experiments to a study of 3- to 5-year-olds who, in addition to one-force events, were asked to predict the path of a ball that was influenced by two forces that were combined to represent force dynamics patterns of 'cause', 'enable' and 'prevent'. The study showed that while the children were successful in their causal reasoning about the one-force events, they attended less to a second force, incorporating it only in the case both forces acted in the same direction. The older they were, the more successful the children became in reasoning about the effects of the second force (George et al., 2019). These experiments indicate that human abstraction and reasoning about physical forces develop with experience over age, even though the general system for perceiving forces as causes is present already at an early age.

A COGNITIVE MODEL OF EVENTS

A Two-Vector Model of Events

The second thesis of this paper is that human causal cognition is structured in terms of *events*. This section argues that mental representations of events exploited in language, physical thinking and planning can be modeled in geometric terms. Several authors (e.g., Talmy, 1988; Croft, 2012; Wolff, 2007, 2008, 2012; Gärdenfors and Warglien, 2012; Gärdenfors et al., 2018) have adopted such a geometric perspective on events. Following earlier work on conceptual spaces (Gärdenfors, 2000, 2014, Gärdenfors and Warglien, 2012; Warglien et al., 2012), I model events as complex structures that involve an action space based on forces and other spaces representing the results of actions.

The two-vector model states that an event is represented in terms of two components – the force of an *action* that generates the event, and the *result* of its application. Both components are represented as vectors in spaces. (In the special case when there is no change, that is, when the result vector is the zero vector, the event is a *state*). The result of an event is modelled as a vector representing the change of properties of the patient before and after the event.

As a simple example of the model, consider the event of Oscar pushing a table. The force vector is generated by the agent Oscar. The result vector is a change in the location of the patient – the table – and thus a change in the properties of the table. The exact result vector depends on the properties of the table, for example its weight as well as other forces in the context, for example, friction. Although typical event representations contain an agent, some need not involve any: for example, events of falling, drowning, dying, growing and raining. The force and result vectors are central, but more vectors and objects may be involved in representations of events as I show below. Following Gärdenfors and Warglien (2012), I put forward the following requirement on the cognitive representation of an event:

The two-vector condition: An event must contain at least two vectors and one object; these vectors are a result vector representing a change in properties of the object and a force vector that causes the change.

The central object of an event will be called the *patient*. If there is an entity generating the force vector, it will be called the *agent* (Wolff, 2007 calls them *force recipient* and *force generator*, respectively). Agents and patients are objects (animate or inanimate) that have different properties. Following my theory of conceptual spaces (Gärdenfors, 2000, 2014), they can be described as vectors of values from property dimensions.

At least two spaces are needed to describe an event, an action space and a result space. The action space can be conceived as a space of forces (or, more generally, force patterns) acting upon some patient, the properties of which are described in the result space. The spaces represent different types of vectors: forces have a different nature than changes in properties.

As the result component of the event represents changes in the properties of the patient, the result space can also be modeled as a vector space. The result vectors typically stand for changes of location or changes of object properties. For example, when Lucy opens the door, the agent Lucy exerts a force vector (action) on the door that leads to a change of the position of the door (result). Or in the event of the storm felling a tree, the force of the wind (action) leads to a change of the direction of the tree (result).

Events are represented not only as single instances, but more generally as event *categories*, for example, throwing a ball. The description of change vectors can be generalized to that of change *vector fields* by associating to each action force vector a result vector, taking into account the (counter-)forces exerted by the patient and other contextual forces. Mathematically, such a mapping from actions to results can be seen as a function from a force vector that is the resulting combination of the action vector and other contextually given forces to a result vector (see Gärdenfors and Warglien (2012) for a more detailed description of the mapping). This mapping is part of the representation of an event category and it contains the central information about causal relations.

The events need not only involve physical forces, but also mental 'forces' can be causal variables (Talmy, 1988; Leslie, 1994). Humans interpret many mental factors (for example commands, threats, insults and persuasive arguments) as forces that can create a change in the physical, cognitive or emotional state of the addressee. For example, Wolff (2007, pp. 19– 22) presents two experiments where a woman intends to cross a street to meet (or to avoid) a man and the directions of a police man in the street crossing acts as an additional 'force' that enables or prevents the woman from reaching her goal. The results show that the subjects interpret the woman's intention as a force and they describe the various scenarios in the same terms as they would use for a situation where only physical forces are involved. In other examples, such as a case of threatening, the resulting change is not physical, but it can still be represented in terms of changes in a conceptual space (assuming that the concept 'person' has a space of emotional states). Wolpert et al. (2003) present an analysis of how this kind of reasoning can be modeled in terms of control theory.

The forces can also be medical, economic or social (Talmy, 1988). For example, in "The aspirin caused his headache to go away," the medicine acts as ca force causing a change in his physical state. And in "The high price offered enabled her to sell her mother's wedding ring," the price acts as a force. A social example is "The pressure from the villagers caused him to mow his lawn, even though he wanted to keep it as a meadow."

I next turn to a more detailed description of the two main components of the model.

Representing Actions

Following Gärdenfors (2007a) [see also Warglien et al. (2012) and Gärdenfors (2014)], I proposed in the previous section that the human cognitive processes extract the forces that generate different kinds of actions. This leads me to the following thesis:

Representation of actions: An action is represented by the pattern of forces that generates it.

The thesis speaks of a pattern of forces since, for most bodily actions, more than one body part is moving. Therefore, multiple force vectors are acting in parallel [this is analogous to Marr and Vaina's (1982) differential equations]. The patterns of forces can be described in the same way as the modeling of shapes in Gärdenfors (2014, Section 6.3). Like shapes, force patterns also exhibit meronomic relations. For example, a bird with short wings flies in a different way than a bird with a large wing span.

In order to investigate the action space, judgements of similarities between actions can be used. The methods for estimating similarities between objects are essentially the same as for objects. The dynamic properties of actions are in focus for such judgments: for example, throwing is more similar to waving than to crawling. A large set of such similarity ratings can serve as data for one of several related statistical techniques, such as multidimensional scaling or principal component analysis that turn similarities into spatial structures. The geometric structure of the action space is largely unknown, except for a few recent studies that are presented below. In line with other domains, it is assumed that the notion of betweenness is meaningful in the action space. This allows me to formulate the following thesis [which is parallel to the thesis about properties in Gärdenfors (2000, 2014)]:

Thesis about action concepts: An action concept is represented as a convex region in the action space.

It is natural to interpret convexity as the assumption that, for any two actions that fall under an action concept, any linear morph between the actions will also belong to the same concept.

Empirical support for the thesis about action concepts involving body movements is presented by Giese and Lappe (2002). Starting from Johansson's (1973) patch-light methods, they edited videos of bodily actions such as walking, running, limping, and marching. Linear combinations of the positions of the joints of the body were created and they then created videos exhibiting morphs of the recorded actions. Subjects who watched the morphed videos were asked to categorize the actions. Giese and Lappe did not explicitly investigate whether the action categories that the subjects created correspond to convex regions. The data they present clearly support convexity.

Another example is Slobin et al. (2014), who investigated how subjects categorized actions shown in 34 video clips of motion events such as walking, running and jumping, The subjects, who were native speakers of English, Polish, Spanish, and Basque, were asked to put a label, as precise as possible, on the action they saw in the clips. Based on the answers a two-dimensional multidimensional scaling solution was calculated. The result indicates that four separated convex regions emerge for each of the languages studied. These regions correspond to walking, running, crawling, and to some non-canonical actions (such as leaping or galloping). Together with similar results from Malt et al. (2014), these results provide support for the thesis about action concepts. However, for human-robot applications, more research concerning the structure of action space is required.

In robotics, the work has mainly dealt with how the *results* of actions can be modelled [e.g., Cangelosi et al. (2008), Lallee et al. (2010), and Demiris and Khadhouri (2006)]. In human-robot interaction, however, it is more important that the robot can categorize human and other actions by the *manner* they are performed. This is called recognition of biological motion (Hemeren, 2008; Gharaee et al., 2017a,b). Categorizing actions is particularly important if the goal of the robot is to understand the intentions behind the actions.

The Causal Mapping

The main reason for introducing the event model is that it is a natural way of capturing how we think about causation: the action *causes* the result. In the literature, most authors analyze the causal relation between the action and the effect as holding between two events (see e.g., Zacks and Tversky, 2001; Casati and Varzi, 2008). In contrast, the model presented here describes causation as a relation *within* an event. Furthermore, the distinction between forces and changes of states also means that the cause and the result, in contrast to traditional theories, are modelled as two different entities.

There are many similarities between the event model presented here and Wolff's (2007, 2008, 2012) dynamics model. His affector vector corresponds to the force vector, his patient vector to the counterforces, and he also includes a result vector. The two models have been developed for slightly different purposes: the two-vector model is presented as a general model of events while the focus of Wolff's model is on causal reasoning. Another difference is that his result vector is of the same kind as the force vectors. In contrast, in the model presented here causes and effects modeled as entities of the different types: they belong to different spaces – causes to the force space and results to change in location space (in the case of movements) or in some property space (color, size, shape, weight, temperature, etc.).

The two-vector model of events has testable consequences. Wolff (2007) presents a study which shows that individuals can perform intuitive addition of force vectors when observing two force simultaneously affect the trajectory of a patient). Michotte's (1963) 'launching' experiments show that how subjects attribute causality in a simulated event involving an object A that hits an object B A depends on the angle of the trajectories of A and B. This shows that subjects judge whether an animation represents one or two events depending on how forces are mapped onto movements. The perception of such a mapping has been shown to be remarkably precise, and to predict the 'causal impression' on the subjects (White, 2012). In these cases, the two-vector model of events predicts well how individuals perceive causal events.

The event model, can handle *what-if* questions, that is, counterfactual reasoning concerning what would have happened if an action would have been different. For example: "If I had dropped the glass on the ceramic floor instead of on the mat, then it would have broken." Such reasoning can be computationally modeled by simulations of various changes in the force and counterforce vectors and using the mapping function and assumed counterforces to predict a result. Simulations use similarity measures and operations for projecting forwards and backwards to understand the causes and consequences. For example, Johnston's (2009) COMIRIT system can be used to integrate commonsense reasoning and the geometric inference of conceptual spaces. COMIRIT establishes a mechanism for assigning 'semantic attachments' to symbols in knowledge representations systems that can be used to automatically construct simulations and utilize machine learning methods. In contrast, probabilistic models of causation, which will be discussed in the subsection Why Probabilistic Models Are Not Suitable, have deep-going problems in handling what-if reasoning (Pearl, 2018).

Similar to counterfactual reasoning, humans often reason in terms of *omissive* causation, that concerns events that do not occur. For example, the fact that a person did not fill in his tax forms, caused that he was fined by the tax authorities. This is a problem for many other models of causation, but the two-vector model can also explain omissive causation [for related solutions see Talmy (1988); Wolff et al. (2010), and Wolff and Thorstad (2017)]. To illustrate how the two-vector model applies in such cases, consider the famous gag in the movie *A Night in Casablanca* where Harpo Marx is leaning against the wall of a house. A policeman comes up to him and says "What do you think you are doing? Holding up the building?" Harpo nods energetically with his typical smile but the policeman chases him away. In the background one sees how the building crashes into the ground. Here, the crash is caused by Harpo's omission of supporting the wall. In the terms of the two-vector model, the force vector from Harpo towards the wall generates a stable state where the wall is in balance despite its counterforces. When Harpo's supporting force is eliminated, the counterforces generate the crash of the house.

THREE CONSTRAINTS ON THE CAUSAL MAPPING

Given our ignorance of the counterforces in a situation and the limited knowledge about the relevant causal relations, it is often very hard to precisely predict the outcome of an action. Still, the qualitative effect of actions can be understood.

When it comes to computational implementations of the two-vector model in a robotic system, the mapping between the force space and the result space is the most central part of the event model. A problem is that externalities, such as friction and other counterforces, make it difficult to determine the result vector, given the force vector. For example, pushing a coffin may result in the coffin moving, other times not; taking a medicine sometimes cures a patient, other times not.

The formal nature of event mappings has been little investigated. Although other theories of events (Talmy, 1988; Croft, 2012, Wolff, 2007, 2008) also build on such a mapping, they do not analyze it. Gärdenfors et al. (2018), however, present an analysis of three general principles for event mappings, that constrain the relation between the force vector and the result vector. All three principles are of a qualitative form, which reflects the qualitative nature of event cognition. They function as ceteris paribus constraints.

As a background for the principles, note that there are three central cognitive processes that depend on mental representations of events: causal thinking, control of action, and learning. These are characterized respectively by three qualitative properties that are central for the corresponding processes: (1) larger forces lead to larger results (this relates to qualitative causal thinking); (2) small changes in forces lead to small changes of the result (this is important for action control); and (3) intermediary results are caused by intermediary forces (this facilitates generalization and categorization of events). Mathematically, these properties correspond respectively to the monotonicity, continuity and convexity preservation of the mapping from actions to results. The motivation for investigating them is that human causal thinking typically satisfies these properties. The three properties thus impose constraints on the mapping from actions to events, something which is crucial when such a mapping is to be learned by a robot.

Larger Forces Lead to Larger Results

A general constraint for qualitative causal thinking is that whenever counterforces and other external factors are kept constant in a given situation, then increasing the force involved in the action will also lead to a larger result (or at least not decrease it). For example, if I push the gas pedal harder in my car, it will run faster.

This constraint captures an important part of our reasoning about how a change of an outcome depends on a change of an action. The constraint makes possible qualitative predictions about the effects of actions. It is a central component in interpreting causality (Hume, 1748/2000; Wolff, 2007, 2008) and in making causal inferences.

The constraint enables qualitative causal inferences. First of all, it makes it possible to draw basic inferences about how changes in causes will lead to changes in effects. For example, since different individuals may react with different intensity to a medicine, it is difficult to predict the size of the effect. One may, however, still make the prediction that increasing the dose of the medicine will increase the effects. Mill (1843) dubbed this form of inference 'the method of concomitant variations'.

Mathematically, this constraint corresponds to the *monotonicity* of the mapping function. A function is said to be monotonous when $f(x) \leq f(y)$, whenever $x \leq y$. This property thus depends on an ordering relation on the forces. As long as all forces act in the same direction such an ordering exist. However, in higher dimensional spaces such an ordering function may not exist.

The constraint that larger forces lead to larger results can also support reverse inference processes. When wanting to identify the relevant causal factors among multiple potential ones, the constraint can provide a powerful selection criterion. For example, the tides have been observed as long as humans have existed, but it was only when the correlations to the moon's position and distance was discovered, taken together with Newton's law of gravitation, that we understood the force vectors causing the tides.

Small Changes in the Force Lead to Small Changes in the Result

When the aim is to change the effect of an action only by a small amount, it can be achieved by applying a correspondingly small change of force. For example, when turning the control for a heater on a stove a little more to the right, one expects the heat also to increase just a little, and not lead to a drastic change that would destroy the food. And when a tennis ball is hit a little harder, it will fly a little faster and further, but not move wide out of the court.

Mathematically, this constraint corresponds to the *continuity* of the mapping function. This can be defined in terms of a nearness relation on the space, which is easily defined for the force space[3].

Central both to human and robotic actions is *motor control*, which in general requires the fine-tuning of an agent's forces (Wolpert and Flanagan, 2001; Stolt et al., 2012). For example, balancing a stick on a finger requires very small adjustments in the neighborhood of the equilibrium position (see e.g., Shiriaev et al., 2007).

While the constraint captures a very general principle of causal thinking, it is not always true that small changes in the force lead to small changes in the result. Sometimes small changes lead to phase transitions. For example, if you are gradually increasing your arm force when bending a wooden stick, there is a point where the stick breaks. At the transition point, a very small change of effort produces a large effect. In more general terms, a discontinuous phase transition occurs when an obstructing counterforce is suddenly overcome, and a drastically different result is achieved.

Intermediate Results Are Caused by Intermediate Forces

Imagine that you are throwing a ball at a basket. You can control the forces of your arms in the throw. If you have tried force *x* and observed that the ball was short of the basket and tried force *y* and observed that the ball went too far, then you presume that a force of a strength between *x* and *y* will lead to an intermediary result.

The third constraint can be formulated as that the causal mapping *f* is convexity preserving: if the force vector *z* is between force vectors *x* and *y*, then the result *f(z)* is between the results *f(x)* and *f(y).* In other *words, intermediate forces lead to intermediate results.* Therefore, this constraint depends on the fact that betweenness is defined for the force and result spaces[4].

This constraint applies to many situations involving bodily movement. A clear example comes from Runesson and Frykholm (1981) who showed subjects patch light movies of a person lifting objects that weighed between two and twenty kilos. The objects themselves were not visible in the movies but only the movement patterns of the person lifting them. In spite of this limited information, the subjects could very accurately predict the weights of the object. The upshot is that the movement patterns were sufficient for the subject to infer the forces that the person lifting the box was applying. The subjects then inferred that intermediary forces corresponded to intermediate weights of the boxes. I am not claiming that the inference is conscious, only that our causal reasoning obeys the constraint.

I have argued that the process of *learning* new concepts requires regions that represent concepts to be convex in order for the process to be efficient (see Gärdenfors, 2000, Ch. 3 and Gärdenfors, 2001). Furthermore, convexity also makes *generalization* efficient since, by interpolation, inferences over whole regions can be made given only a limited number of observations. Finally, feedback control mechanisms also require that the mapping from actions to results preserves convexity (e.g., Shiriaev et al., 2007).

It should be noted that generalization in psychology has focused on generalizing from a particular data point (for example Shepard, 1987). However, generalizing by interpolation between data points is at least as important. Given that convexity is satisfied, it is sufficient to know the mappings from two force vectors to two result vectors to know what lies between them. Thus, convexity helps to predict unspecified properties of the event[5].

To sum up this section, the three qualitative constraints do not uniquely determine the mapping from causes to results, but they add rich structure to it. The constraints make it possible to draw robust inferences even if counter-forces and other contextual factors are unknown. In this way, the constraints considerably strengthen human causal thinking. It is therefore recommendable that robotic systems for causal reasoning also obey these constrains.

The three constraints have been presented here as part of the two-vector event model presented in Section 3. Because of their general nature, however, they can also be applied to other models such as the force dynamics of Talmy (1988), the dynamic model of Wolff (2007, 2008) and the event representations in Croft (2012).

IMPLEMENTING THE EVENT MODEL AND THE CAUSAL MAPPING IN ROBOTS

The core of the two-vector model of events consists of the mapping between the force space and the result space. In this section, I present some considerations on how the mapping – and how it is learned – may be implemented in a robotic system.

Learning the Event Mapping: Computational Aspects

As a simple but illustrative case, I will take Wolff's (2007, 2008) studies of how people evaluate causes and effects of how controlling the speed and direction of the motor of a boat will affect its trajectory. A complicating factor is that, apart from the resistance of the water, there is an unpredictable wind that acts as a counterforce. In this causal web, the physics of the situation allows a system to learn the unknown variables. Firstly, in situations without wind the effects of the speed and direction of the force vector of the motor can be learned (and it will be a linear mapping as long as friction is constant), since

[3] The precise definition is: A mapping *f*: $X \rightarrow Y$ between topological spaces is called *continuous* if the pre-image under *f* of any open subset of *Y* [denoted $f^{-1}(Y)$] is an open subset of *X*. I should be noted that any metric induces a nearness relation.

[4] Again, a metric induces a betweenness relation. If S is a space with a metric d, then *z* in *S* lies *between x, y* in *S* if $d(x,y) = d(x,z) + d(z,y)$.

[5] Gärdenfors et al. (2018) argue that these constraints are central for the 'working model' of an event (Zacks et al., 2007; Radvansky and Zacks, 2014).

the friction vector is always in the opposite direction of the force vector. Secondly, once this mapping is learned, one can simulate situations where there is a wind, and by adding the friction and wind counterforce vectors, the system can learn to identify a motor force vector that will result in the desired effect. There are several ways of computationally implementing such a learning system by using traditional physical modeling or by using some form of neural network. I will not go into details here.

Other situations will not admit such a principled learning procedure. In many cases there may be unknown counterforces and other factors that make the mapping non-linear and dependent on several external variables. However, by letting the system experience a number of varied data points, approximations of a mapping function can be calculated. When a so far unobserved result vector is desired, interpolations of force vectors resulting in similar effects can be used to generate a new force vector that, because of the three constraints of the mapping function, result in an approximate result vector. For the implementation of learning situations of this kind, many methods from control theory can be employed (see e.g., Ardakani et al., 2019).

Even in situations where the forces are non-physical, similar methods can be used to learn the event mapping. For example, Wolpert et al. (2003) explore the computational parallels between motor control, on the one hand, and action observation, imitation, and social interaction, on the other (see also Gärdenfors, 2007b). They argue that motor commands that generate bodily actions can be extended to social actions directed towards other people. In this extension, the changes in the state of my body correspond to changes of the state of mind of another person.

Another field of learning that is required for robotic reasoning about causation and for communicating, for example in a planning situation, is *action categorization*. Representations of actions in terms of conceptual spaces, such as those proposed by, for example, Chella et al. (2001), Gärdenfors (2014), and Gharaee et al. (2017a,b), provide a potentially fruitful method for implementations. Simulating an action and then using the event mapping that has been learned to predict a result vector, can then be used to generate plans and to reason about complex situations. In this way, simulations can provide the robotic system the power to imagine events that is needed to understand the physical, social and, eventually, the emotional world we live in.

The event structure has not yet been implemented in any concrete system. However, a cognitively motivated architecture for holistic AI systems, including robotic ones, that integrates machine learning and knowledge representation has been proposed in Gärdenfors et al. (2019). The central idea of the proposal is to use ‘event boards’ representing components of events as an analogy to blackboards that formed the backbone in some earlier AI systems. The event components that are placed on the board are represented by vectors in conceptual spaces rather than in symbolic structures that has been used in previous systems. A control level that is added to the event board includes an attention mechanism that decides which processes are run.

Why Probabilistic Models Are Not Suitable

Within computer science, Bayesian models or Bayesian nets are popular statistical tools since they require minimal prior knowledge (see Waldmann and Hagmayer, 2013 for a presentation). For example, ‘constraint-based algorithms’ allow the derivation of causal structures on the basis of the pattern of statistical dependencies of a set of variables (see e.g., Pearl, 2000). Another way of learning causal structure is to formulate the problem in terms of Bayesian inferences. For such a learning mechanism, the learning system (for example, a robot) must determine the probability of a causal structure given the available data. There also exist proposals for hybrid systems combining Bayesian models with more traditional models (Waldmann and Mayrhofer, 2016).

There are, however, some problems connected with probabilistic models (Wolff, 2007; Waldmann and Hagmayer, 2013), in particular when it comes to implementations on robotic systems. In experimental studies, subjects have had difficulties in extracting causal relations based on covariation data even though these experiments typically present a small number of variables (Steyvers et al., 2003). For humans, a single instance of a causal connection is sufficient to pick up a causal relation and it would be desirable that a robotic system has a similar capacity. Such a rapid process is difficult to capture in a probabilistic model. According to the model presented here, the forces that generate an action are essential for causal inferences and such forces are, in general, inaccessible to probabilistic approaches. In brief, Bayesian processes are computationally not suitable for implementations in robotic systems.

The implausibility of domain-general algorithms of structure induction has led Waldmann (1996) to propose the view that people generally use prior hypothetical knowledge about the structure of causal models to guide learning in a top down fashion, so called knowledge-based causal induction. In line with this, also Waldmann and Hagmayer (2013) argue that causal cognition of people cannot be encompassed by the Bayesian formalism. For these reasons, I do not consider the Bayesian approach to be a viable alternative for robotic systems[6]. Furthermore, the use of the general principles of monotonicity, continuity and convexity makes much of Bayesian reasoning unnecessary.

CONCLUSION

In this article, I have argued for two theses. The first thesis is that human causal cognition (in contrast to that of non-human animals) build on understanding the *forces* that are involved in an action that leads to a result. The second thesis is that humans think about causality in terms of *events*. I have presented the

[6] Pearl’s (2000) model requires that the causal structure of the variables is provided in advance.

two-vector model of events that is based on conceptual spaces and shown that it captures several aspects of human causal reasoning.

I have argued that Bayesian models are not suitable for representing causal structures, in particular not the event structures that have been presented here. The two-vector model of events generate new types of problems that must be solved in order to create robotic systems capable of causal reasoning. The main problem is to devise methods for learning appropriate mappings from actions to results, that is, from causes to effects.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Ardakani, M. M. G., Olofsson, B., Robertsson, A., and Johansson, R. (2019). Model predictive control for real-time point-to-point trajectory generation. *IEEE Trans. Autom. Sci. Eng.* 16, 972–983. doi: 10.1109/tase.2018.2882764

Boesch, C., Bombjaková, D., Boyette, A., and Meier, A. (2017). Technical intelligence and culture: nut cracking in humans and chimpanzees. *Am. J. Phys. Anthropol.* 163, 339–355. doi: 10.1002/ajpa.23211

Cacchione, T., Call, J., and Zingg, R. (2009). Gravity and solidity in four great ape species (*Gorilla gorilla*, *Pongo pygmaeus*, *Pan troglodytes*, *Pan paniscus*): vertical and horizontal variations of the table task. *J. Comp. Psychol.* 123, 168–180. doi: 10.1037/a0013580

Call, J. (2010). "Trapping the minds of apes: causal knowledge and inferential reasoning about object-object interactions," in *The Mind of the Chimpanzee: Ecological and Experimental Perspectives*, eds E. V. Lonsdorf, S. R. Ross, T. Matsuzawa, and J. Goodall (Chicago, IL: Chicago University Press), 75–86.

Cangelosi, A., Metta, G., Sagerer, G., Nofi, S., Nehaniv, C., Fischer, K., et al. (2008). "The iTalk project: Integration and transfer of action and language knowledge in robots," in *Proceedings of Third ACM/IEEE International Conference on Human Robot Interaction*, Vol. 2, Amsterdam, 167–179. doi: 10.1111/tops.12099

Casati, R., and Varzi, A. (2008). "Event concepts," in *Understanding Events: From Perception to Action New*, eds T. F. Shipley and J. Zacks (New York, NY: Oxford University Press), 31–54.

Chella, A., Gaglio, S., and Pirrone, R. (2001). Conceptual representations of actions for autonomous robots. *Rob. Auton. Syst.* 899, 1–13.

Croft, W. (2012). *Verbs: Aspect and Causal Structure*. Oxford: Oxford University Press.

Davidson, D. (1967). "The logical form of action sentences," in *The Logic of Decision and Action*, ed. N. Rescher (Pittsburgh, PA: University of Pittsburgh Press), 81–95.

Demiris, Y., and Khadhouri, B. (2006). Hierarchical attentive multiple models for execution and recognition of actions. *Rob. Auton. Syst.* 54, 361–369. doi: 10.1016/j.robot.2006.02.003

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2001). Concept learning: a geometric model. *Proc. Aristotelian Soc.* 101, 163–183. doi: 10.1111/j.0066-7372.2003.00026.x

Gärdenfors, P. (2003). *How Homo Became Sapiens: On the Evolution of Thinking*. Oxford: Oxford University Press.

Gärdenfors, P. (2007a). "Evolutionary and developmental aspects of intersubjectivity," in *Consciousness Transitions: Phylogenetic, Ontogenetic and Physiological Aspects*, eds H. Liljenström and P. Århem (Amsterdam: Elsevier), 281–305. doi: 10.1016/b978-044452977-0/50013-9

Gärdenfors, P. (2007b). Mindreading and control theory. *Eur. Rev.* 15, 223–240. doi: 10.1017/S1062798707000233

Gärdenfors, P. (2014). *Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: MIT Press.

Gärdenfors, P., Jost, J., and Warglien, M. (2018). From actions to events: three constraints on event mappings. *Front. Psychol.* 9:1391. doi: 10.3389/fpsyg.2018.01391

Gärdenfors, P., and Lombard, M. (submitted). Technology made us understand abstract causality.

Gärdenfors, P., and Lombard, M. (2018). Causal cognition, force dynamics and early hunting technologies. *Front. Psychol.* 9:87. doi: 10.3389/fpsyg.2018.00087

Gärdenfors, P., and Warglien, M. (2012). Using conceptual spaces to model actions and events. *J. Semant.* 29, 487–519. doi: 10.1093/jos/ffs007

Gärdenfors, P., Williams, M.-A., Johnston, B., Billingsley, R., Vitale, J., Peppas, P., et al. (2019). "Event boards as tools for holistic AI," in *Proceedings of the 6th International Workshop on Artificial Intelligence and Cognition, CEUR Workshop Proceedings*, Vol. 2418, eds A. Chella, I. Infantino, and A. Lieto (Palermo: University of Technology Sydney), 1–10.

George, N. R., Göksun, T., Hirsh-Pasek, K., and Golinkoff, R. M. (2019). Any way the wind blows: children's inferences about force and motion events. *J. Exp. Child Psychol.* 177, 119–131. doi: 10.1016/j.jecp.2018.08.002

Gharaee, Z., Gärdenfors, P., and Johnsson, M. (2017b). Online recognition of actions involving objects. *Biol. Inspired Cogn. Arch.* 22, 10–19. doi: 10.1016/j.bica.2017.09.007

Gharaee, Z., Gärdenfors, P., and Johnsson, M. (2017a). First and second order dynamics in a hierarchical SOM system for action recognition. *Appl. Soft Comp.* 59, 574–585. doi: 10.1016/j.asoc.2017.06.007

Giese, M., Thornton, I., and Edelman, S. (2008). Metrics of the perception of body movement. *J. Vis.* 8, 1–18. doi: 10.1167/8.9.13

Giese, M. A., and Lappe, M. (2002). Measurement of generalization fields for the recognition of biological motion. *Vis. Res.* 42, 1847–1858. doi: 10.1016/s0042-6989(02)00093-7

Göksun, T., George, N. R., Hirsh−Pasek, K., and Golinkoff, R. M. (2013). Forces and motion: how young children understand causal events. *Child Dev.* 84, 1285–1295. doi: 10.1111/cdev.12035

Hanus, D., and Call, J. (2008). Chimpanzees infer the location of a reward on the basis of the effect of its weight. *Curr. Biol.* 18, R370–R372.

Hemeren, P. (2008). *Mind in Action*. Lund: Lund University Cognitive Studies, 140.

Hume, D. (1748/2000). *An Enquiry Concerning Human Understanding*. Oxford: Clarendon Press.

James, W. (1890). *The Principles of Psychology*, Vol. 1. London: Macmillan.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/bf03212378

Johnson-Frey, S. H. (2003). What's so special about human tool use? *Neuron* 39, 201–204. doi: 10.1016/s0896-6273(03)00424-0

Johnston, B. (2009). *Practical Artificial Commonsense*. Ph.D. thesis, University of Technology, Sydney.

Köhler, W. (1917). *The Mentality of Apes*. Mitchan: Penguin Books.

Lallee, S., Madden, C., Hoen, M., and Dominey, P. F. (2010). Linking language with embodied and teleological representations of action for humanoid cognition. *Front. Neurorob.* 4:8. doi: 10.3389/fnbot.2010.00008

Leslie, A. M. (1994). "ToMM, ToBy, and agency: core architecture and domain specificity," in *Mapping the Mind: Domain Specificity in Cognition and Culture*, eds L. A. Hirschfeld and S. A. Gelman (New York, NY: Cambridge University Press), 139–148.

Leslie, A. M. (1995). "A theory of agency," in *Causal Cognition: A Multidisciplinary Debate*, eds D. Sperber, D. Premack, and A. J. Premack (Oxford: Oxford University Press), 121–141.

Leslie, A. M., and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition* 25, 265–288. doi: 10.1016/s0010-0277(87)80006-9

Lombard, M., and Gärdenfors, P. (2017). Tracking the evolution of causal cognition in humans. *J. Anthropol. Sci.* 95, 1–16. doi: 10.4436/JASS.95006

Malt, B., Ameel, E., Imai, M., Gennari, S., Saji, N. M., and Majid, A. (2014). Human locomotion in languages: constraints on moving and meaning. *Mem. Lang.* 74, 107–123. doi: 10.1016/j.jml.2013.08.003

Marr, D., and Vaina, L. (1982). Representation and recognition of the movements of shapes. *Proc. R. Soc. Lond. B* 214, 501–524.

Martin-Ordas, G., Call, J., and Colmenares, F. (2008). Tubes, tables and traps: great apes solve two functionally equivalent trap tasks but show no evidence of transfer across tasks. *Anim. Cogn.* 11, 423–430. doi: 10.1007/s10071-007-0132-1

Michotte, A. (1963). *The Perception of Causality*. New York, NY: Methuen.

Mill, J. S. (1843). *A System of Logic*. London: John W. Parker.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge, MA: MIT Press.

Pearl, J. (2018). Theoretical impediments to machine learning: with seven sparks from the causal revolution. *arXiv* [Preprint]. arXiv:1801.04016vi.

Penn, D. C., and Povinelli, D. J. (2007). Causal cognition in human and nonhuman animals: a comparative, critical review. *Annu. Rev. Psychol.* 58, 97–118. doi: 10.1146/annurev.psych.58.110405.085555

Povinelli, D. (2000). *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works*. Oxford: Oxford University Press.

Povinelli, D., and Penn, D. C. (2011). "Through a floppy tool darkly: toward a conceptual overthrow of animal alchemy," in *Tool Use and Causal Cognition*, eds T. McCormack, C. Hoerl, and S. Butterfill (Oxford: Oxford University Press), 69–97.

Radvansky, G. A., and Zacks, J. M. (2014). *Event Cognition*. Oxford: Oxford University Press.

Runesson, S. (1994). "Perception of biological motion: the ksd-principle and the implications of a distal versus proximal approach," in *Perceiving Events and Objects*, eds G. Jansson, S. S. Bergström, and W. Epstein (Hillsdale, NJ: Lewrence Erlbaum associates), 383–405.

Runesson, S., and Frykholm, G. (1981). Visual perception of lifted weights. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 733–740. doi: 10.1037/0096-1523.7.4.733

Seed, A., and Call, J. (2009). "Causal knowledge for events and objects in animals," in *Rational Animals, Irrational Humans*, eds S. Watanabe, A. P. Blaisdell, L. Huber, and A. Young (Minato: Keio University Press), 173–188.

Seed, A., Hanus, D., and Call, J. (2011). "Causal knowledge in corvids, primates and children: more than meets the eye?," in *Tool Use and Causal Cognition*, eds T. McCormack, C. Hoerl, and S. Butterfill (Oxford: Oxford University Press), 89–110. doi: 10.1093/acprof:oso/9780199571154.003.0005

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323. doi: 10.1126/science.3629243

Shiriaev, A. S., Freidovich, L. B., Robertsson, A., Johansson, R., and Sandberg, A. (2007). Virtual-holonomic-constraints-based design of stable oscillations of Furuta pendulum: Theory and experiments. *IEEE Trans. Rob.* 23, 827–832. doi: 10.1109/tro.2007.900597

Slobin, D. I., Ibarretxe-Antuñano, I., Kopecka, A., and Majid, A. (2014). Manners of human gait: a crosslinguistic event-naming study. *Cogn. Linguist.* 25, 701–741.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., and Blum, B. (2003). Inferring causal networks from observations and interventions. *Cogn. Sci.* 27, 453–489. doi: 10.1016/S0364-0213(03)00010-7

Stolt, A., Linderoth, M., Robertsson, A., and Johansson, R. (2012). Adaptation of force control parameters in robotic assembly. *IFAC Proc. Vol.* 45, 561–566. doi: 10.3182/20120905-3-hr-2030.00033

Talmy, L. (1988). Force dynamics in language and cognition. *Cogn. Sci.* 12, 49–100. doi: 10.1523/JNEUROSCI.0447-17.2017

Tomonaga, M., Imura, T., Mizuno, Y., and Tanaka, M. (2007). Gravity bias in young and adult chimpanzees (Pan troglodytes): tests with a modified opaque−tubes task. *Dev. Sci.* 10, 411–421. doi: 10.1111/j.1467-7687.2007.00594.x

Waldmann, M. R. (1996). "Knowledge-based causal induction," in *The Psychology of Learning and Motivation, Vol. 34: Causal Learning*, eds D. R. Shanks, K. J. Holyoak, and D. L. Medin (San Diego, CA: Academic Press), 47–88.

Waldmann, M. R., and Hagmayer, Y. (2013). "Causal reasoning. To appear," in *Oxford Handbook of Cognitive Psychology*, ed. D. Reisberg (New York, NY: Oxford University Press).

Waldmann, M. R., and Mayrhofer, R. (2016). "Hybrid causal representations," in *The Psychology of Learning and Motivation*, Vol. 65, ed. B. Ross (New York, NY: Academic Press).85-127

Wang, W., Crompton, R. H., Carey, T. S., Günther, M. M., Li, Y., Savage, R., et al. (2004). Comparison of inverse-dynamics musculo-skeletal models of al 288-1 australopithecus afarensis and knm-wt 15000 homo ergaster to modern humans, with implications for the evolution of bipedalism. *J. Hum. Evol.* 47, 453–478.

Warglien, M., Gärdenfors, P., and Westera, M. (2012). Event structure, conceptual spaces and the semantics of verbs. *Theor. Linguist.* 38, 159–193.

White, P. A. (2012). Visual impressions of causality: effects of manipulating the direction of the target object's motion in a collision event. *Vis. Cogn.* 20, 121–142.

Wolff, P. (2007). Representing causation. *J. Exp. Psychol. Gen.* 13, 82–111.

Wolff, P. (2008). "Dynamics and the perception of causal events," in *Understanding Events: How Humans See, Represent, and Act on Eve*nts, eds S. Thomas and J. Zacks (Oxford: Oxford University Press), 555–587.

Wolff, P. (2012). Representing verbs with force vectors. *Theor. Linguist.* 38, 237–248.

Wolff, P., Barbey, A. K., and Hausknecht, M. (2010). For want of a nail: how absences cause events. *J. Exp. Psychol. Gen.* 139, 191–221. doi: 10.1037/a0018129

Wolff, P., and Shepard, J. (2013). "Causation, touch, and the perception of force," in *The Psychology of Learning and Motivation*, Vol. 58, ed. B. H. Ross (New York, NY: Academic Press), 167–202.

Wolff, P., and Thorstad, R. (2017). "Force dynamics," in *The Oxford Handbook of Causal Reasoning*, ed. M. R. Waldmann (New York, NY: Oxford University Press), 147–167.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 593–602.

Wolpert, D. M., and Flanagan, J. R. (2001). Motor prediction. *Curr. Biol.* 11, R729–R732.

Woodward, J. (2011). "A philosopher looks at tool use and causal understanding," in *Tool Use and Causal Cognition*, eds T. McCormack, C. Hoerl, and S. Butterfill (Oxford: Oxford University Press), 18–50.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychol. Bull.* 133, 273–293.

Zacks, J. M., and Tversky, B. (2001). Event structures in perception and conception. *Psychol. Bull.* 127, 3–21.

Conflict of Interest: 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.

ORIGINAL RESEARCH
published: 06 May 2020
doi: 10.3389/fpsyg.2020.00872

Inferring Unseen Causes: Developmental and Evolutionary Origins

Zeynep Civelek*, **Josep Call and Amanda M. Seed**

School of Psychology and Neuroscience, University of St Andrews, St Andrews, United Kingdom

Human adults can infer unseen causes because they represent the events around them in terms of their underlying causal mechanisms. It has been argued that young preschoolers can also make causal inferences from an early age, but whether or not non-human apes can go beyond associative learning when exploiting causality is controversial. However, much of the developmental research to date has focused on fully-perceivable causal relations or highlighted the existence of a causal relationship verbally and these were found to scaffold young children's abilities. We examined inferences about unseen causes in children and chimpanzees in the absence of linguistic cues. Children (*N* = 129, aged 3–6 years) and zoo-living chimpanzees (*N* = 11, aged 7–41 years) were presented with an event in which a reward was dropped through an opaque forked-tube into one of two cups. An auditory cue signaled which of the cups contained the reward. In the causal condition, the cue followed the dropping event, making it plausible that the sound was caused by the reward falling into the cup; and in the arbitrary condition, the cue preceded the dropping event, making the relation arbitrary. By 4-years of age, children performed better in the causal condition than the arbitrary one, suggesting that they engaged in reasoning. A follow-up experiment ruled out a simpler associative learning explanation. Chimpanzees and 3-year-olds performed at chance in both conditions. These groups' performance did not improve in a simplified version of the task involving shaken boxes; however, the use of causal language helped 3-year-olds. The failure of chimpanzees could reflect limitations in reasoning about unseen causes or a more general difficulty with auditory discrimination learning.

Keywords: causal reasoning, hidden causes, temporal order, pre-schoolers, chimpanzees

Edited by:
Andrea Bender,
University of Bergen, Norway

Reviewed by:
Alex Taylor,
The University of Auckland,
New Zealand
Katja Liebal,
Freie Universität Berlin, Germany
Selma Dündar-Coecke,
University College London,
United Kingdom

****Correspondence:***
Zeynep Civelek
zc8@st-andrews.ac.uk

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

Received: *12 November 2019*
Accepted: *07 April 2020*
Published: *06 May 2020*

Citation:
Civelek Z, Call J and Seed AM
(2020) Inferring Unseen Causes:
Developmental and Evolutionary
Origins. Front. Psychol. 11:872.
doi: 10.3389/fpsyg.2020.00872

INTRODUCTION

In life and also in science, much of the evidence we get for causal relations is indirect. We can infer the existence and nature of a cause for an event despite not witnessing it directly: if it is hidden from our perspective, or if it is not perceivable by the senses. Our inferences can range from identifying the cause of a crashing sound coming from the kitchen (the wooden cutting board or the metal pot falling on the floor) to the causes of global warming (anthropogenic impact on the greenhouse effect). But how do we do this? Bullock et al. (1982) suggest that we use the principles of determinism, priority and mechanism: We assume that there is a causal structure to the world (i.e., that events typically have causes); that these structures are unidirectional (i.e., causes come before their effects) and that events are underpinned by a causal mechanism of some kind. Using these principles, and our prior knowledge with regards to specific relations, we can work our way from effects to detect likely causes. This is an extraordinary ability that frees us from relying on what can be directly perceived, allows us to make predictions about the future, and intervene to bring about desirable outcomes.

However, we can also learn regular covariations in spatiotemporal contiguity, which allow us to exploit a causal pattern even if we do not theorize about the generative mechanism (Shanks and Dickinson, 1987). If two events occur repeatedly under close spatiotemporal proximity, we form associative links between them. Later when one of the cues occur, the other can be predicted without any reference to the causal mechanism involved, indeed, without any explicit awareness of the relationship at all (Reber, 1989). Conversely, we can learn a great deal about unseen causal relations without any direct experience: from others' explicit testimony or implicit linguistic cues to causality (Harris and Koenig, 2006; Gelman, 2009). We may even learn about causal relations we may not have learnt otherwise (e.g., "The gravitational attraction of the moon causes tides"). These three alternative routes to exploiting causal relations in the world (association, theory-building and testimony) are not mutually exclusive, as adults we make us of all of them, and they interact in important ways.

What are the origins of these abilities in human development and over human evolution? There is good evidence that statistical or associative learning is present early in infancy (Aslin et al., 1998; Kirkham et al., 2002), and that this ability is shared with a great many other species. It is similarly uncontroversial that learning from testimony is a route available to children once they learn language, and unique to our species. However, when it comes to going beyond the data to reason about causal mechanisms there is more controversy both in developmental and comparative psychology (Penn and Povinelli, 2007; Bonawitz et al., 2010; Seed et al., 2011). Some researchers have suggested that humans have a natural tendency to explain the events they observe in terms of causal theories from very early in life (Bullock et al., 1982; Gopnik and Wellman, 2012). If this is the case, it is plausible that we share this ability with our closest primate relatives, and possibly other species (Seed et al., 2011; Völter and Call, 2017). Alternatively, others contend that causal thinking in early childhood might not be well-characterized by the notion of "theories all the way down" (Carey and Spelke, 1996). Instead children's thinking about causation may only approximate scientific thinking later in development, due in part to input from others with the development of language. If this is the case, we may not expect to find causal reasoning in non-human primates. Penn and Povinelli (2007) have argued that there is no evidence non-human animals represent causality as such.

While tackling these questions empirically, one issue common to the comparative and developmental literature concerns distinguishing causal reasoning (based on representations of causal mechanism) from associative learning (making predictions in the absence of these representations), since events that are causally linked tend to co-occur. From a developmental perspective alone, a second issue concerns teasing apart the role of causal language and reasoning since children can use both to solve causal problems. We have two aims in this paper: (1) to further explore children's inferences about unseen causes in the absence of linguistic cues to causality, and (2) to use the same paradigm to explore this ability in our closest relatives, chimpanzees.

There is substantial research suggesting that preschool children take unseen causal relations into account when explaining natural phenomena such as light (Bullock et al., 1982), wind (Shultz, 1982), electricity (Buchanan and Sobel, 2011), and contamination by germs (Legare et al., 2009). However, it is difficult to isolate the route to causal knowledge in cases that involve familiar events such as these. Children may have extensive prior experience with lights and blowing candles which may lead to forming associative links or may have been explicitly taught by adults about how "germs cause disease." Indeed, younger preschoolers who supposedly did not have extensive experience with wires and electricity, failed to reason about these relations and made decisions based on covariation information instead (Buchanan and Sobel, 2011). They were only able to solve the problem when it involved more familiar batteries. Although it is possible that experience leads to extracting abstract causal information, it may also lead to learning arbitrary associations (e.g., when there are batteries inside, the toy works).

A way to address this issue has been to present preschoolers with novel and arbitrary causal structures. As adults and scientists, when the evidence we get does not fit with our prior knowledge or expectations, we infer unseen causes or confounding variables. In order to test if children reasoned in the same way, children were first trained on a novel causal structure (e.g., puppets moving in a certain way), and then saw evidence that was inconsistent with their training (Gopnik et al., 2004; Schulz and Sommerville, 2006; Schulz et al., 2008). When children were asked to make predictions about the cause of this inconsistent event, they were more likely to say that an unseen cause (i.e., "something else") was responsible. Children also displayed an ability to imagine the effect of a hidden cause in a series of experiments by Siegel et al. (2014). They were able to select boxes to shake that would yield unambiguous data (e.g., if their task was to locate a hard object, they chose to pair it with a soft object rather than another hard object). However, in these studies the existence of a cause and the possibility that it might be unseen was provided in the framing of the task by the experimenter so the children did not have to infer it from the evidence alone. For instance, the experimenter asked "Why are the puppets moving together? Is it X, Y or something else?"

Overall, the evidence suggests that by 4 years of age children can successfully detect the presence of an unseen cause and make inferences about their nature; but the potential impact of others' verbal testimony on their abilities has not been explored to date. Gelman (2009) argued that children are not "lone scientists": they get much needed input from adults around them. Linguistic framing can help children to specify a causal relation by testifying that the covariations they see are indeed causal; and the use of same wording can point to the commonalities between an observed action and agent's action (as in intervention studies: "The block makes it go. Can you make it go?"). Indeed, there is accumulating evidence that the use of causal framing can impact children's propensity to make causal inferences from directly perceived and indirect evidence (Sobel and Sommerville, 2009; Bonawitz et al., 2010; Butler and Markman, 2012; Lane and Shafto, 2017).

One possibility is that verbal framing merely highlights the problem for children: making the task more sensitive to their theory construction ability by reducing peripheral demands such as the need to focus attention (Sobel and Sommerville, 2009). Another possibility is that without the verbal framing younger children are yet to develop some of the fundamental cognitive components needed to construct a causal explanation from evidence alone. The difficulty with using never-seen-before causal relationships is that some training or explanation is necessary for children to have the required background information to make inferences. While the nature of the instructions have been varied, they are rarely excluded. The verbal framing may simplify the task for older children, equally, it may make the test unsuitable for younger children such as 2–3 year-olds if they lack sufficient verbal ability to follow the instructions. We therefore designed a paradigm with minimal language requirements to explore this issue. We also intended to use this paradigm to make comparisons between children and non-human primates. This line of evidence could be very informative in establishing the degree to which human scientific thinking is grounded in skills we share with our closest relatives, or is rather a skill that requires cultural input over development to emerge, and verbal input to elicit in younger children.

Whether or not our closest relatives, chimpanzees, engage in causal reasoning is a controversial issue in comparative psychology. Some authors propose that causal reasoning is a uniquely human ability; and chimpanzees either learn associatively or they rely on generalizations based on the surface appearance of objects alone to solve problems (Penn and Povinelli, 2007; Penn et al., 2008; Bonawitz et al., 2010). Limitations in performance in some tasks designed to probe the causal reasoning abilities of great apes would seem to support this interpretation (Köhler, 1925; Limongelli et al., 1995; Povinelli, 2000; Call, 2007). In contrast to Penn and Povinelli (2007), Seed et al. (2011) proposed that non-human great apes can make use of causal information from events happening around them if the testing situation does not overload other cognitive resources. It could be shown that they did not rely solely on the available sensory information to learn associations. However, it has been a challenge to decisively distinguish associative learning from causal reasoning.

One of the most promising ways to resolve this issue has been to compare how non-human primates (and other animals, such as corvids and dogs) make inferences about the location of food in two contexts, either: (a) the evidence is caused by the food or (b) the evidence co-varies with the presence of food but the relation is arbitrary (reviewed in Seed and Mayer, 2017; Völter and Call, 2017). Great apes successfully used indirect evidence to locate food in a number of studies: in the form of auditory cues coming from shaken cups (Call, 2004), the visible effect of weight (Hanus and Call, 2008); and visible traces or trails (Völter and Call, 2014). In the critical comparison conditions, in which the relationship between a similar cue and the food location was arbitrary rather than causal, apes did not find the food (for example, if the experimenter played the recording of the rattling sound over the baited cup, Call, 2004). Taken together these studies imply that apes are capable of causal reasoning about unseen causes.

However, the comparability of the arbitrary conditions to the causal ones were criticized. For example, Penn and Povinelli (2007) point out that the "recorded sound" control of the shaken cups study was not identical to the sound the shaken cup made. They further argued that the results could still be explained by associative learning if subjects had used the combination of shaking motion and rattling sound as a discriminative cue for locating food. Overall, the comparability of the experimental and control conditions in terms of different feedback (e.g., auditory) poses a challenge for distinguishing causal reasoning from associative learning.

The task presented in this study was designed to address some of the empirical challenges raised above by reducing verbal requirements and implementing robust controls for associative learning. In the "causal condition," a ball containing a reward was dropped into a forked tube, and could be found in one of two cups at the bottom. After the ball was dropped, participants heard either a *ding* or a *clack* sound. After a few trials, subjects were expected to learn that when they heard a *ding*, the ball would be in one cup and when it was a *clack*, the ball would be in the other one. If subjects succeed in this condition, it might mean that they reasoned about the underlying causal structure (the ball hitting the different boxes caused different sounds) or that they simply associated the sound with the side (if *ding*, choose right). In order to distinguish between these two possibilities, in the "arbitrary condition" the order of events was reversed: participants first heard a *ding* or a *clack* sound, and then the ball was dropped into the forked tube. Although the sounds were still predictive of the location of the ball (if *ding*, choose right), the relationship was now arbitrary. Critically, the two conditions were equivalent from an associative learning perspective since the stimuli involved in both conditions were exactly the same and the only difference was the order of events. However, if participants reason about unseen causes, they are expected to do better in the "causal condition" where there is a plausible causal structure than in the "arbitrary condition."

In previous studies, we have found such differences between causal and arbitrary conditions in children between the ages of 3 and 5, when dealing with directly perceivable events such as choosing an appropriate tool or an unobstructed path for extracting a reward (Mayer et al., 2014; Seed and Call, 2014). However, such performance differences are not apparent in older children, probably because 6-year olds are capable of interpreting arbitrary cues as symbolic communication to solve a problem (DeLoache, 2004; Seed et al., 2011; Mayer et al., 2014). We therefore focused on the 3–6-year-olds in this study. By 3-years of age children expect causes to precede their effects (Bullock and Gelman, 1979; Rankin and McCormack, 2013) so we predicted that by this age children should perform at above chance levels in the causal condition if they reasoned causally, and by 6-years they should be above chance in both conditions.

EXPERIMENT 1: CHILDREN

Methods

Participants

Three-to-six-year-old children (N = 129) were tested in different locations in Scotland. There were 65 children in the causal condition and 64 in the arbitrary condition. Age and sex were split roughly equally in the two conditions (**Table 1**). Twenty-three additional children that were tested were excluded from the study due to experimenter or apparatus error (7), parental interference (3), discovery of the trick about the box (4) and refusal to complete the task (9). All the children studies reported in this paper were ethically approved by University of St Andrews Teaching and Research Ethics Committee and informed consent were taken from parents/guardians.

Materials

Transparent training box

The training apparatus was a forked chute made from clear acrylic (**Figure 1**). The middle singular channel (30 × 6.5 × 5cm) was forked into two channels. Directly at the bottom of the channels there were two white acrylic boxes (2.5 cm apart). The channels were mounted on a white acrylic back panel (30 × 49 cm); and a base panel (25 × 30 cm) to stand. They contained pegs that were 7.5 cm apart from each other on both sides. The pegs were designed to slow down the fall of the ball and to make sounds so that subjects could easily follow the ball's trajectory. A peg positioned right above the fork could be moved to the either side from behind the back panel. It enabled the experimenter to control which side the ball would fall in a trial.

TABLE 1 | Age, sex, and mean/median performances of children in Experiments 1, 2, 4, and 5.

	N (females)	Mean age	Mean/Median performance	SD
Experiment 1				
3-year-olds				
Causal	16 (8)	3.6	0.45	0.50
Arbitrary	16 (8)	3.4	0.50	0.50
4-year-olds				
Causal	16 (8)	4.5	**0.62**	0.48
Arbitrary	16 (7)	4.4	0.50	0.50
5-year-olds				
Causal	16 (8)	5.4	0.55	0.49
Arbitrary	16 (8)	5.4	0.48	0.50
6-year-olds				
Causal	17 (8)	6.4	**0.60**	0.49
Arbitrary	16 (8)	6.3	**0.60**	0.49
Experiment 2				
4–5-year-olds				
Causal	20 (9)	4.7	**0.61**	0.49
Arbitrary	20 (10)	4.6	**0.62**	0.49
Experiment 4				
3-year-olds	16 (8)	3.5	0.54	0.50
4-year-olds	16 (9)	4.5	**0.69**	0.46
5-year-olds	16 (8)	5.4	**0.83**	0.38
Experiment 5				
3-year-olds	28 (14)	3.7	**0.67**	0.47

Numbers in bold represent means that are significantly different from chance (p < 0.05) according to a Wilcoxon signed rank test (Experiment 1) and one sample t-tests (Experiments 2, 4, and 5).

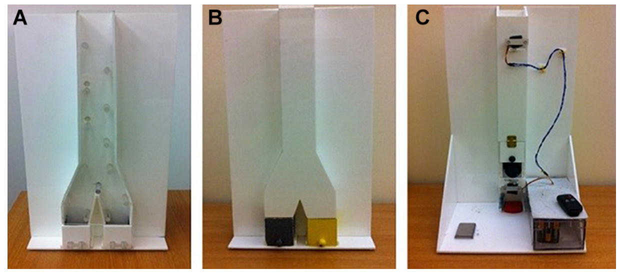


FIGURE 1 | Transparent training box **(A)**, opaque testing box **(B),** and the back of the opaque testing box **(C)** used in Experiments 1 and 3.

Opaque testing box

The testing apparatus had the same measurements as the training box but the channels were opaque. The boxes at the bottom of the channels were spray-painted, one yellow and one gray, using Plastikote stone-textured paint (**Figure 1**). In the testing apparatus, the back panel concealed two additional elements which, unbeknownst to the participant, controlled the falling of the ball through the apparatus and the production of the sound cues.

First, there was a middle singular channel (30 × 10 cm) into which the dropped ball would fall, hitting pegs along the way, and land noiselessly on a piece of foam. Below this channel was a shorter one (6.5 cm) in which a second ball was held and could be released onto a noise-making block (wooden or metal). This block could be exchanged by the experimenter depending on the trial. These two components were combined through the action of two small motors which controlled the rotation of small plastic supports that held the two balls in place. When the motors were switched on by a remote, the plastic supports would rotate, releasing the two balls according to a precise timing. The two buttons on the remote controlled the order in which the motors would activate. In the *causal condition*, the motor at the top would operate first and let the ball dropped into the apparatus by the experimenter, go down the channel hitting the pegs, and then the motor at the bottom would release the second ball to fall onto the metal/wooden piece positioned by the experimenter. The intended illusion was that the ball had fallen down the channel into one of the two boxes and made a distinct sound. In the *arbitrary condition*, the activation of the motors was reversed. The second support moved first to release the ball on the metal/wooden piece, and then the experimenter dropped the ball in time for the first support to rotate and let the ball fall down the channel with the pegs. The time interval between the activation of the two motors copied the actual time it would take the ball to fall in reality and was the same in both conditions. In the causal condition it appeared as a single event sequence. The electronic card that controlled the motors was concealed in a box behind the apparatus (**Figure 1**). The reason for creating the illusion rather than using a real event sequence was that: (1) no local sound cues were given to locate the ball; and (2) the order of the cues could be reversed in the arbitrary condition while keeping everything else about the stimuli exactly the same.

The balls were made of thermoplastic (1.60 cm in diameter) and contained a hole in the middle where the reward could be put.

Procedure

Training phase

The experiment started with the transparent training box. The experimenter introduced the task saying; "In this game, I will put a sticker in the ball and then I will drop the ball from here (the top opening). It will roll down to one of these boxes (points to the boxes at the bottom). If you find the ball, you will win the sticker. Ready?" The experimenter then dropped the ball and the child could watch the entire trajectory of the ball until it came to rest, hidden, in one of the boxes. The child then pointed to or opened the box that she/he thought the ball was in. Once the child made a choice, the other box was also opened to show the content. Transparent training ended after five consecutive successes or ten trials in total.

Test phase

After the training, the experimenter said "This game was too easy for you! Shall we make it more fun?" and brought out the opaque testing box. Then introduced the task to the children; "The game is the same. I will put a sticker in the ball and drop the ball from here. If you can find the ball, you will win the sticker. You cannot see inside the box anymore, but there is still a way to find the ball in the correct box! Do you want to try?" Before each trial, the experimenter prepared the apparatus behind a barrier by putting a ball with a sticker inside into one of the boxes at the bottom, placing another ball on the support attached to the motor just above the metal/wood piece and holding another in her hand for the child to see. The metal and wood pieces were interchanged in between trials and the remote that controlled the events rested behind the apparatus.

In the *causal condition*, the experimenter pressed the causal-order button on the remote while dropping the ball. From the participant's perspective, they would see the experimenter drop a ball into the apparatus, follow the trajectory of the fall due to the pegs inside the middle channel and then hear a metallic or wooden sound.

In the *arbitrary condition*, the experimenter pressed the arbitrary-order button on the remote. The participant would first hear a metallic/wooden sound and then see the experimenter drop the ball into the apparatus and follow the trajectory of the fall due to the pegs.

If the child found the ball, the experimenter said "Well done! You won a sticker!" removed the other box to show that it was empty and prepared for the next trial. If the child did not find the ball, the experimenter said "Oh no! It was here (opening the other box). Let's do it again!" In total children got 20 testing trials which lasted about 15 min. For a given participant the position of the yellow and gray boxes at the end of the channels stayed the same over the 20 trials (e.g., the yellow box on the left was associated with a *ding*, and the gray box in the right was associated with a *clack*), but between subjects the pairing of the color of the box and the sound were randomized. The ball was placed in each box 10 times in a random fashion but never in the same box more than twice in succession.

Open ended question

At the end of the task, the experimenter asked children; "How did you decide which box to choose?" If children did not reply, the experimenter elaborated "Sometimes the ball was in the gray one and sometimes in the yellow one. How did you know where the ball was?" Other than 15 missing explanations (first 10 participants were not asked because it was not initially planned in the study design and 5 other participants had to leave immediately after testing), all children responded to the question.

Scoring and Analysis

The first choice of the subjects was scored as their response in all of the experiments. All trials were scored live by the experimenter as correct or incorrect and were also videotaped. A second examiner coded 20% of the videos for reliability, *Kappa* = 0.97 (95% CI [0.95, 0.99], $p < 0.001$. The mistakes that were found by the second coder were corrected and all the videos were recoded from the video once again to check for other potential mistakes (none were found). The data for this study can be found at **Supplementary Table S2**.

We specified generalized linear mixed models (GLMM; Baayen, 2008) with binomial error structure and logit link function using the function glmer of the R-package lme4 (Bates et al., 2015) for all of our analyses in this paper. In Experiment 1, our full model comprised of condition (causal/arbitrary), age, and their interaction; trial number, and sex as fixed effects. Subject ID and the side of the boxes were included as random effects. In order to keep type-1 error rates at the nominal level of 5%, we included random slopes of trial number within subject ID, but left out the correlation parameters between random intercepts and random slopes terms (Schielzeth and Forstmeier, 2009; Barr et al., 2013). We compared the full model to a null model which included only the random effects using a likelihood ratio test.

The model stability was assessed by excluding individual cases one at a time and comparing the estimates with those derived from a model with the full data set. The model was stable with regards to the fixed effects. We checked whether the variability was greater than expected (overdispersion) and found that it was not an issue with regards to the final model (dispersion parameter: 0.95). Finally, variance inflation factors (VIF) were calculated using the function vif of the R-package car and it did not indicate collinearity to be an issue.

The data was not normally distributed so non-parametric Wilcoxon signed-rank tests were used to examine whether children's performance was significantly different from chance level ($p = 0.05$) in different conditions and age groups. Children who chose one side 16 or more times were counted as side biased according to a two-tailed binomial test ($p = 0.004$). Chi-square tests were used to explore the relationship between side bias, condition and age.

Children's responses to the open-ended questions were categorized into five types of explanations (N = 114) using the relevant categories from Legare et al. (2010). The first category, "No explanation," consisted of children who could/did not provide a verbal strategy (e.g., pointed to the boxes, said "yellow/gray one"). The second category was "Don't know," which consisted of children who said they did not know how to find the ball and they were just guessing. The third category was "Non-causal strategies" that referred to a solution based on a non-causal feature or pattern (e.g., the ball alternated right-left-right-left, "because of the colors"). The fourth category was "Causal explanations that were wrong" (e.g., "I followed the noises into the boxes," "The box wiggled a bit when the ball fell into it"). And the last category was "Referring to different sounds/materials" which showed an understanding of the true causal structure (e.g., "They made two different sounds"). A second examiner categorized children's answers into these five different types of explanations. There was a high agreement between the two coders, *Kappa* = 0.86 [95% CI, 0.80, 0.93], $p < 0.001$ and it rose to *Kappa* = 0.96 [95% CI, 0.92, 0.99], $p < 0.001$ after further discussions. The disagreements were due to some responses that could be categorized either as category one or two (e.g., "Don't know" and points to the boxes). We decided to include them in "no explanation" category as they were mostly pointing gestures. Only when children explicitly stated that they were just guessing, we included them in "Don't know" category. The relationship between verbal explanations, age and condition was explored using chi-square tests.

Results

Training

All children except for one 5 and three 3-year-olds passed the transparent training within 5 consecutive trials. Two of these children needed 6 and the other two needed 8 trials to complete the transparent training.

Test

The full model comprising of the interaction of age and condition, sex and trial number as fixed effects fit the data better than the null model which lacked these fixed effects [$\chi^2(9) = 30.91, p < 0.001$]. We found that there was a significant condition and age interaction [$\chi^2(3) = 8.71$, $p < 0.05$] and a significant effect of trial number [$\chi^2(1) = 6.99$, $p < 0.01$]. There was no effect of sex [$\chi^2(1) = 3.17$, $p = 0.075$] (**Supplementary Table S1**).

Comparisons of children's performance in different conditions across age groups showed that there was no significant difference between performance in the causal and arbitrary conditions for 3- and 6-year olds (Mann–Whitney U-Test for 3-year-olds: $U = 93$, $N_{\text{causal}} = 16$, $N_{\text{arbitrary}} = 16$, $p = 0.348$; 6-year-olds: $U = 133$, $N_{\text{causal}} = 17$, $N_{\text{arbitrary}} = 16$, $p = 0.921$). Three-year-olds performed at chance level in both causal (Median: 0.45, Wilcoxon signed-ranks test: $T^+ = 72$, $N = 15$, $p = 0.513$) and arbitrary conditions (Median: 0.5, $T^+ = 40$, $N = 11$, $p = 0.562$); 6-year-olds were above chance in both causal (Median: 0.6, $T^+ = 127$, $N = 17$, $p < 0.05$) and arbitrary conditions (Median: 0.6, $T^+ = 92.5$, $N = 14$, $p < 0.01$). Four-year-olds performed significantly better ($U = 64.5$, $N_{\text{causal}} = 16$, $N_{\text{arbitrary}} = 16$, $p < 0.05$) and above chance levels in causal condition (Median: 0.62, $T^+ = 98.5$, $N = 14$, $p < 0.01$) as opposed to chance level performance in arbitrary condition (Median: 0.5, $T^+ = 39$, $N = 12$, $p = 1$). Five-year-olds showed a similar trend for better performance compared to chance in the causal condition (Median: 0.55, $T^+ = 61$, $N = 12$, $p = 0.08$) than in the arbitrary condition (Median: 0.48, $T^+ = 66$, $N = 15$, $p = 0.751$); however this difference was not significant ($U = 93$, $N_{\text{causal}} = 16$, $N_{\text{arbitrary}} = 16$, $p = 0.191$). **Figure 2** shows the average performance of each age group in causal and arbitrary conditions. An effect of learning as evidenced by the significant effect of trial number on performance was found. This was expected given that subjects had no way of solving the task in their first trial.

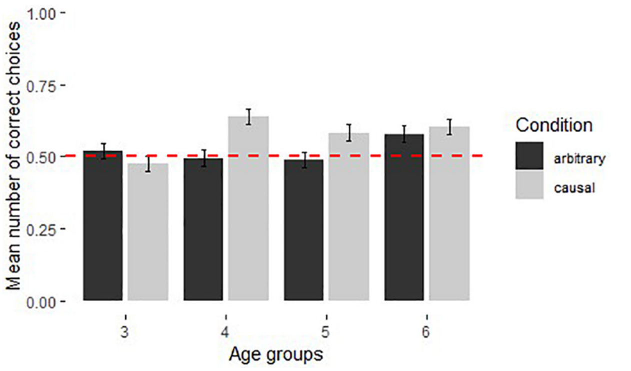


FIGURE 2 | Performance of children in the causal and arbitrary conditions in Experiment 1 (N = 129, see **Table 1** for age group information and means). Dotted line shows chance level performance (p = 0.05), error bars represent *SE*.

There was no significant relationship between condition and side-bias [$\chi^2(1) = 0.73$, $p = 0.39$], however, there was a significant relationship between age and side bias [$\chi^2(3) = 16.77$, $p < 0.001$]. Three-year-olds were more likely to be side biased than other age groups.

Open Ended Question

Table 2 summarizes the percentages of children's responses to the question "How did you decide which box to choose?" in each age group across two conditions. For a more robust analysis using chi-square, "no explanation" and "don't know" categories; and "non-causal strategies" and "wrong causal explanations"

TABLE 2 | Percentage of children who gave the following explanations in response to the question "How did you know where the ball was?" in Experiment 1 (N = 114).

	Causal condition				Arbitrary condition			
Explanations	3 yo	4 yo	5 yo	6 yo	3 yo	4 yo	5 yo	6 yo
No idea	90%	58.3%	43%	31%	81%	67%	33%	34%
Wrong idea	10%	8.3%	21%	25%	19%	33%	60%	41%
Correct explanation	0%	33.3%	36%	44%	0%	0%	7%	25%

categories were lumped to result in three explanation categories in total: "no idea," "wrong idea," and "correct explanation." According to the chi-square analysis there was not a significant relationship between 3-year-old children's explanations and the condition they were in [$\chi^2(1) = 0.36$, $p = 0.55$]. In both conditions, a high percentage of 3-year-olds had "no idea" about how to find the ball, a minority gave a wrong explanation and there were no children who could provide the correct explanation. There was a significant relationship between 4-year-olds' explanations and condition [$\chi^2(2) = 6.95$, $p < 0.05$]. Although the majority of 4-year-olds were in the "no idea" category in both conditions, 33.3% could provide the correct explanation in the causal condition whereas none did in the arbitrary condition. Interestingly, there was a higher percentage of children in the arbitrary condition who gave a "wrong idea" explanation compared to those in the causal condition. The relationship between explanations and condition were marginally significant for 5-year-olds [$\chi^2(2) = 5.73$, $p = 0.057$]. "No idea" responses were comparable in both conditions, however, there were more 5-year-olds in the arbitrary condition who referred to wrong explanations than in causal condition and there were more children in causal condition that referred to the "correct explanation" than in the arbitrary condition. There was a significant relationship between 6-year-olds' explanations and the condition they were in [$\chi^2(2) = 6.51$, $p < 0.05$]. The pattern was similar to 5-year-olds. More children referred to wrong explanations in the arbitrary condition compared to the causal condition and more 6-year-olds in the causal condition came up with an explanation based on different sounds than in the arbitrary condition.

Finally we explored whether children's reports matched with their performance. The performance of the 16 children who referred to different sounds/materials in the causal condition was compared with the performance of an age-matched group in the causal condition who gave other explanations. The model comprising of the fixed effects of explanations (correct/incorrect), trial number and sex fit the data better than the null model without the fixed effects [$\chi^2(3) = 29.47$, $p < 0.001$] (**Supplementary Table S2**). There was a significant effect of explanation type on children's performance [$\chi^2(1) = 24.90$, $p < 0.001$]. Children who gave the correct explanations performed better than their peers who gave incorrect explanations [Mean difference = 0.25, 95% CI [0.15, 0.36], $t(15) = 5.24$, $p < 0.001$]. Moreover, children who gave correct explanations performed above chance levels [$M = 0.77$, 95% CI [0.69, 0.85], $t(15) = 7.17$, $p < 0.001$], whereas those who gave incorrect explanations were at chance [$M = 0.52$, 95% CI [0.47, 0.56], $t(15) = 0.76$, $p = 0.46$].

Discussion

When the sound cues were consistent with a causal structure, by 4-years of age children used the discriminatory sound cue to locate the ball, whereas 3-year-olds failed. When the cues were not consistent with a causal structure, 4–5-year-olds did not use these same sounds to find the ball; and performed worse than they did in the causal condition. This difference was significant for 4-year-olds but not for 5-year-olds. These results suggested that children went beyond the immediately available cues to imagine their likely unseen causes. The explanations children provided about how they found the ball matched the results of the main task. More children referred to different sounds/materials when there was a plausible causal structure than when the relation was arbitrary. In addition, the children who referred to different sounds outperformed their peers who gave different explanations for their choice.

However, one could argue that the temporal proximity between the distinct sound cue (metal/wood) and the outcome (choice of one box) was smaller in the causal condition: when the order of events was "falling" (filler) sound, metal/wood sound, choice, than in the arbitrary condition, when the order was metal/wood sound, filler sound, and then choice. And since associations are more easily formed between temporally proximate events (Barnet et al., 1991; Miller and Barnet, 1993), and even brief delays have been shown to result in a reduction of causality judgments (Michotte, 1963; Shanks et al., 1989), these could explain the better performance in the causal condition compared to the arbitrary condition. In Experiment 2 we tested this alternative explanation.

Six-year-olds performed equally well in both conditions. Their successful performance in the arbitrary condition might have resulted from the ability to treat arbitrary cues as symbols to solve a problem (DeLoache, 2004; Seed et al., 2011; Mayer et al., 2014). On the other hand, 3-year-olds did not pass either condition in this study, they were unable to provide a verbal explanation about how they found the ball and were more likely to be side-biased.

One possibility for the failure of 3-year-olds could be that unlike older children, they cannot, or do not spontaneously, imagine unseen causes. However, other explanations are possible too, such as the necessity to remember the cues which, being auditory, are transitory, and map them to one of the two boxes which do not look to be made of the materials evoked by the sounds. In Experiment 4 we simplified the task by using boxes that were visibly made of metal and wood, to examine whether or not this task would be easier. In Experiment 3, we tested chimpanzees, and planned to titrate the level of difficulty based on our initial results with the task described above.

EXPERIMENT 2: FOLLOW-UP WITH ARBITRARY SOUNDS

In this experiment, we tested whether better performance in the causal condition as opposed to the arbitrary condition in Experiment 1 could be due to temporal proximity of the sound cues and the outcome. Children were asked to locate a sticker in one of the two boxes based on recorded sounds which were similar either to the causal (filler, wood/metal) or the arbitrary order (wood/metal, filler) of the Experiment 1. Would children perform better when the discriminatory cue was more proximate to the choice, than when it was followed by a filler sound? If this was the case, then it would raise concerns that the differences between the causal and arbitrary conditions in Experiment 1 could be due to temporal proximity rather than causal plausibility. However, if children detected the causal structure, we did not expect to find differences between conditions when all cues, regardless of the order, were arbitrarily related to the outcome.

Methods

Participants

A new group of 40 4–5-year-old children were tested. Half of them participated in the *filler discriminatory* condition and the other half participated in the *discriminatory, filler* condition. Age and sex were split roughly equally in the two conditions (**Table 1**). Two additional children that were tested were excluded from the study due to refusal to complete the task.

Materials

The yellow and gray boxes at the bottom of the channels in Experiment 1 were used. The boxes were covered with lids so that children could not see inside. A barrier (52 × 31 cm) concealed the hiding event. Two sounds that lasted about 1 second were recorded and played back to the children from the experimenter's phone. The sounds were amplified with a speaker.

Procedure

Test

The experimenter introduced the task saying; "In this game, I will hide a sticker in one of these boxes. You won't see where it goes but there is a way to find the sticker in the correct box. I will give you the clue using my phone! If you point to the correct box, you will win the sticker. Ready?" The experimenter then hid the sticker behind the barrier. Upon removing the barrier, she said "Now, pay attention!" and played the recorded sound from her phone. In the *filler, discriminatory* condition, the children heard the filler sound followed by a metal/wood sound at the end and in the *discriminatory, filler* condition, they heard the metal/wood sound followed by the filler sound. The experimenter asked "Where do you think is the sticker?" and the child pointed to or opened the box that she/he thought the sticker was in. Once the child made a choice, the other box was also opened to show the content.

In total children got 20 trials which lasted about 10 min. The side of the boxes was randomized across subjects. The sticker was placed in each box ten times in a random fashion but never in the same box more than twice in succession.

Open ended question

At the end of the task, the experimenter asked; "How did you decide which box to choose?" If children did not reply, the experimenter elaborated "Sometimes the ball was in the gray one and sometimes in the yellow one. How did you know where the ball was?" A second examiner categorized children's answers into these five different types of explanations, *Kappa* = 0.92 [95% CI, 0.85, 0.96], $p < 0.001$.

Scoring and Analysis

A second examiner coded 20% of the videos for reliability, *Kappa* = 0.98 [95% CI, 0.97, 0.98], $p < 0.001$. The full model consisted of condition, trial number and sex as fixed effects; ID and the side of the boxes as random effects. We also included random slopes of trial number within ID, but left out the correlation parameters between random intercepts and random slopes terms (Schielzeth and Forstmeier, 2009; Barr et al., 2013). This full model was compared to a null model which included only the random effects using a likelihood ratio test.

The model was stable, overdispersion was not an issue with regards to the full model (dispersion parameter: 0.86) and there was no multicollinearity. We used one-sample t-tests to examine performance different from chance level (p = 0.05) in the two conditions. Finally, the relationship between verbal explanations and condition was explored using chi-square tests.

Results

Test

The full model was not significantly different from the null model [$\chi^2(3) = 0.38$, $p = 0.945$]. None of the predictors had a significant influence on performance (**Supplementary Table S3**). Children performed above chance in both the causal [M = 0.61, 95% CI [0.52, 0.71], $t(19) = 2.48$, $p < 0.05$] and the arbitrary sound orders [M = 0.62, 95% CI [0.53, 0.70], $t(19) = 2.87$, $p < 0.01$] (**Figure 3**).

Open Ended Question

There were 2 missing data points so the analysis was conducted on data from 38 children. There was not a significant relationship between children's explanations and the condition they were in [$\chi^2(2) = 3.03$, $p = 0.22$]. Overall, there were 13 children who were in the "no idea" group; 14 children in "wrong explanations" and 11 children who gave the correct explanation. Only the children who were in the correct explanation group performed above chance level [$t(10) = 7.24$, $p < 0.001$].

Discussion

When the sound cues to locate the reward was completely arbitrary, the order children heard them did not influence

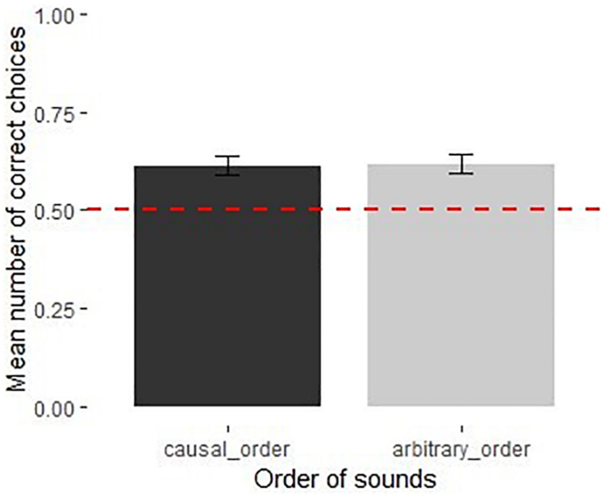


FIGURE 3 | Performance of children in the causal and arbitrary recorded sound orders in Experiment 2 (*N* = 40). Dotted line shows chance level performance (*p* = 0.05), error bars represent *SE*.

their performance and there was no relationship between their verbal explanations and the condition they were in. Therefore, better performance in the causal than the arbitrary condition in Experiment 1 could not simply be explained based on the temporal proximity of the sound cue and the outcome as the associative accounts would suggest. Indeed, Buehner and May (2002, 2003) have also challenged the necessity of temporal proximity for causal judgments by showing that it was the knowledge of the causal structure that influenced participants' judgments.

Overall, this experiment provided further evidence to support our interpretation that by 4-years of age children were able to use indirect evidence to detect unseen causes based on data alone. In Experiment 3 we explore chimpanzees' abilities to detect unseen causes.

EXPERIMENT 3: CHIMPANZEES

The experiment with chimpanzees consisted of two phases. In the first phase, we planned to test 6 subjects in the causal condition as described in Experiment 1. If subjects passed the causal condition, in the second phase, we planned to test a further 6 chimpanzees on the arbitrary condition. However, if they did not pass the causal condition, in the second phase we planned to simplify the task by replacing the yellow and gray boxes at the bottom with metal and wooden boxes (familiar boxes). With this manipulation the subjects would receive additional visual feedback with the conspicuously metal and wooden boxes that could help them match the sounds and materials more easily.

Methods

Participants

Chimpanzees housed at the Wolfgang Köhler Primate Research Center, Leipzig Zoo (Germany), were selected by convenience sampling. Six chimpanzees participated in the first phase: causal condition with unfamiliar boxes. Because none of these individuals passed the task at above chance levels, in the second phase, 3 of these experienced chimpanzees and 3 additional naïve subjects were assigned to the "Familiar boxes, causal condition" and the other 3 experienced and 3 additional naïve subjects were assigned to the "Familiar boxes, arbitrary condition." One subject in the "familiar boxes arbitrary condition" stopped approaching the mesh for testing after a few sessions, so she was dropped from the study, leaving 11 subjects in total who participated in the second phase (**Table 3**). Subjects lived in two groups of 6 and 19 individuals and had access to indoor and outdoor enclosures. They were tested individually in their sleeping rooms and were not deprived of food or water at any time. Testing days were consecutive as much as possible. If a subject did not choose to participate, testing for this individual was canceled for that day. Research was conducted in accordance with the regulations of the University of St Andrews' Animal Welfare and Ethics Committee (AWEC), Max Planck Institute for Evolutionary Anthropology and Zoo Leipzig.

TABLE 3 | The name, age, sex, rearing history and information about experiment participation of chimpanzees (*N* = 11).

Name	Age	Sex	Rearing history	Participation (condition)
Hope	26	f	Nursery	Unfamiliar (causal), familiar (causal)
Kofi	11	m	Mother	Unfamiliar (causal), familiar (causal)
Fraukje	41	f	Nursery	Unfamiliar (causal), familiar (causal)
Bangolo	7	m	Mother	Familiar (causal)
Sandra	24	f	Mother	Familiar (causal)
Lobo	13	m	Mother	Familiar (causal)
Tai	14	f	Mother	Unfamiliar (causal), familiar (arbitrary)
Dorien	36	f	Nursery	Unfamiliar (causal), familiar (arbitrary)
Riet	39	f	Nursery	Unfamiliar (causal), familiar (arbitrary)
Lome	15	m	Mother	Familiar (Arbitrary)
Frodo	23	m	Mother	Familiar (arbitrary)

Materials

Transparent training and opaque testing boxes

Exact replicas of the apparatuses described in Experiment 1 were used to test chimpanzees in the unfamiliar boxes causal condition (**Figure 1**). The apparatus was placed on a sliding table (78 × 38 cm) which was fixed to the sides of the mesh panel (78 × 55 cm). A second opaque screen was placed behind the mesh panel to block the view of the subject in between trials.

In the second phase with familiar boxes, the yellow and gray boxes were replaced with boxes of the same size made of wood and metal (**Figure 4**).

Sound making training boxes

Chimpanzees in Leipzig Zoo had objects made of different materials in their outdoor and indoor enclosures (i.e., automatic metal feeders, tree logs, plastic buckets) and occasionally may hear the noises they make when they are hit/dropped. However, in comparison to children we assumed their exposure to metal and wooden materials would be limited. Therefore, we prepared two sound-making training boxes: the "metal box" (6 × 5.5 × 6 cm) made from stainless steel and the "wooden box" (8 × 7 × 7 cm) made from ply-wood. In both boxes, there was a thermoplastic ball (1.30 cm in diameter); and there was a hole (1.25 cm in diameter) on one side of the box. The boxes also

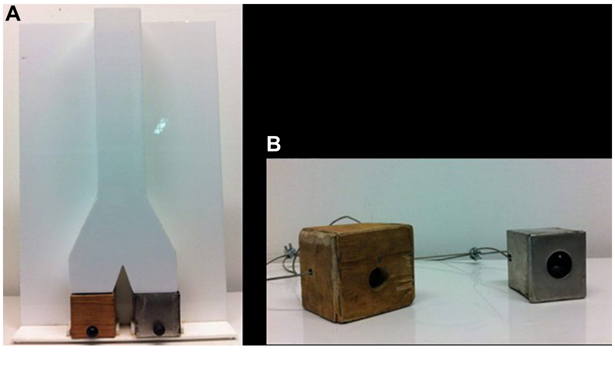


FIGURE 4 | Testing box used in familiar boxes (wooden and metal) conditions **(A)** and sound-making training boxes **(B)** in Experiment 3.

contained peanuts which could be shaken free, an action which caused the ball to rattle inside the box and make sounds. The boxes were passed to the chimpanzees through a movable feeder that was adjacent to the mesh panel they were tested. A steel wire passing through each sound-making box secured them to the steel feeder. Therefore, the subjects could play with the boxes but could not take them away (**Figure 4**).

Procedure

Training phase

All subjects completed the transparent training phase before moving on to the testing. The experimenter placed the transparent apparatus on the sliding table, put a food reward (dates, peanuts based on subjects' preference) in the ball and when the subject was sitting in front of the mesh, dropped the ball from the top opening. The subjects were highly motivated to find the high value food rewards, and were familiar with the experimental setup where they tried to locate rewards in cups/boxes/apparatuses. When the experimenter pushed the sliding table toward the mesh, the subjects could point to one of the boxes at the bottom. If the subject chose correctly, the experimenter gave the reward to the subject and took out the other box to show that it was empty. If the subject pointed to the wrong box, the experimenter first showed the empty box and then showed the content of the other box and put the food reward back into the bucket. If the subject pointed to an irrelevant location or the choice was ambiguous, the experimenter pulled the sliding table back, tapped on both boxes at the same time and pushed the table forward again. When a trial was over, the opaque screen was put behind the mesh. Chimpanzees received 10 trials per session and training continued until the subjects selected the correct side 16 out of 20 trials or more (a binomial test was run to calculate p-value, $p = 0.004$).

Once a subject passed the transparent training the subjects also received the sound-making boxes training. The experimenter put shelled peanuts in full view of the subject into one of the boxes and passed it to the subject using the steel feeder. When the subject shook the boxes, the ball hit the walls of the box making metal/wooden sounds and the peanuts came out through the hole. Once the subject was done with one box, the experimenter replaced it with the other box. Half of the subjects got the metal box first and the wooden second and the other half did the reverse order. They got sound-making boxes training at the beginning of each testing session.

Test phase

The *unfamiliar boxes causal condition* was the same as described above in Experiment 1. Chimpanzees got 10 trials per session and testing ended when a subject selected the correct side 16 out of 20 times or more or until 10 sessions were completed. The side of the yellow/gray boxes at the end of the channels were randomized across subjects. The ball was placed in each box 5 times in a random fashion but never in the same box more than twice in succession.

The procedure for the *familiar boxes conditions* with wooden and metal boxes at the bottom were the same as the unfamiliar boxes.

Scoring and Analysis

A second examiner coded 20% of the videos for reliability, *Kappa* = 0.81 [95% CI, 0.75, 0.87], $p < 0.001$. In the *unfamiliar boxes causal condition* the full model comprised of age, sex, session and trial numbers as fixed effects and ID and the side of the boxes as random effects. We included random slopes of trial and session numbers within ID, but left out the correlation parameters between random intercepts and random slopes terms. The full and null model comparison was done using a likelihood ratio test. In order to explore performance in this condition against chance level ($p = 0.05$) we used a one-sample t-test. In the *familiar boxes conditions*, same analyses methods were used with the addition of condition (causal/arbitrary) and experience (experienced/naïve) to fixed effects.

Both models for unfamiliar and familiar boxes were stable and there were no issues with regards to overdispersion (*dispersion parameter: 1.01* for both*)*, however, multicollinearity was an issue for the predictors, age and sex. Therefore, sex was removed from the models.

Results

Unfamiliar Materials

Transparent training

All chimpanzees except for one reached the criterion in the transparent training within two sessions which was the minimum amount. This subject needed an extra session to reach the criterion.

Testing

None of the subjects reached the criterion in the unfamiliar boxes causal condition; therefore, all subjects received 10 sessions (see **Figure 5**). The full model was not significantly different from the null model [likelihood ratio test: $\chi^2(3) = 1.43$, $p = 0.698$] (**Supplementary Table S4**). Furthermore, they were at chance level overall as a group [$M = 0.50$, 95% CI [0.45, 0.56], $t(5) = 0.15$, $p = 0.885$]. All individuals except for one were side biased.

Since none of the subjects in unfamiliar boxes condition passed the task, we moved on to the familiar boxes.

Familiar Materials

Transparent training

Five naïve subjects got the transparent training before moving on to the testing sessions. They reached the criterion within two sessions.

Testing

The model including condition, experience level, age, session and trial numbers was not significantly different from the null model [$\chi^2(5) = 2.74$, $p = 0.741$]. There was no significant difference between performances in the causal and arbitrary conditions nor between the performances of experienced and naïve individuals (**Supplementary Table S5**). Subjects in both conditions performed at chance level; familiar boxes causal [$M = 0.52$, 95% CI [0.46, 0.57], $t(5) = 0.76$, $p = 0.480$] and familiar boxes arbitrary [$M = 0.53$, 95% CI [0.50, 0.56], $t(4) = 2.67$, $p = 0.06$] (**Figure 6**). One subject reached the criterion in the causal familiar condition in the last session ($M = 0.62$, $SD = 0.16$);

whereas none of the subjects in the arbitrary condition passed the task. All individuals except for one in the familiar boxes arbitrary condition were side biased. There was no significant relationship between condition and side bias [$\chi^2(1) = 1.32$, $p = 0.251$].

Discussion

Chimpanzees were at chance level in both the causal and arbitrary conditions and there were no significant differences between them. Negative results are difficult to interpret, and while the results, for chimpanzees as for 3-year-olds, could speak to limitations in spontaneously imagining an unseen cause, there are other explanations that could account for their failure. For example, the requirement to integrate knowledge about how the channels worked (i.e., the ball can land in any one of the boxes randomly) with the sounds of different materials could have been challenging. In a recent study, chimpanzees did not spontaneously cover the two exits of a similar forked chute, suggesting that this kind of event might be difficult for them to anticipate (Suddendorf et al., 2017). Therefore, in Experiment 4, we simplified the task further by removing the channels completely, and simply requiring subjects to infer where the ball was based on the sound of one of the boxes being shaken.

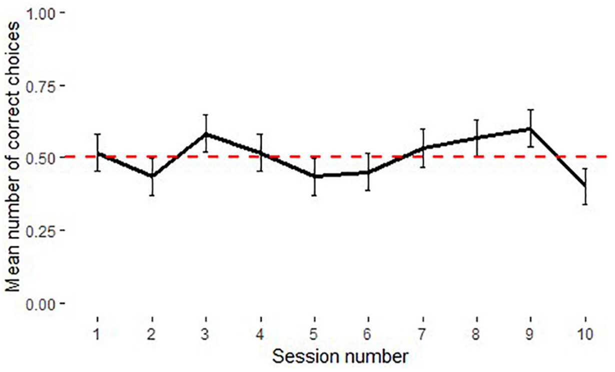


FIGURE 5 | Performance of chimpanzees in the unfamiliar boxes causal condition across sessions in Experiment 3 Phase 1 (*N* = 6). Dotted line shows chance level performance (*p* = 0.05), error bars represent *SE*.

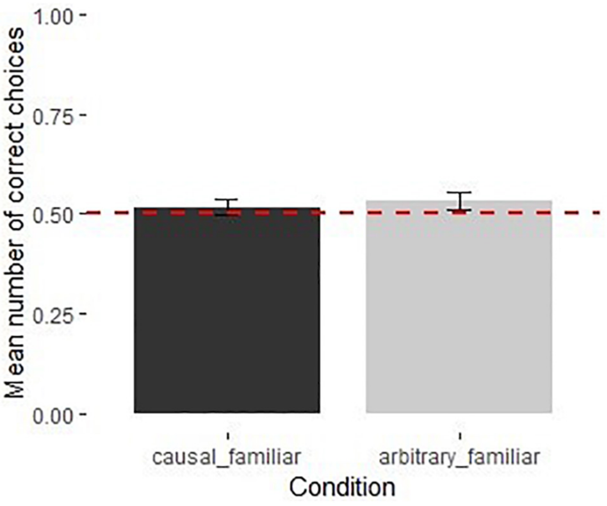


FIGURE 6 | Performance of chimpanzees in the familiar boxes causal and arbitrary conditions in Experiment 3 Phase 2 (*N* = 11). Dotted line shows chance level performance (*p* = 0.05), error bars represent *SE*.

EXPERIMENT 4: SHAKEN BOXES

We aimed to see if children and chimpanzees could infer the location of a food reward in one of two boxes (made of wood and metal) based on the different sounds made when a ball was shaken in one of the boxes behind a barrier. We predicted that 4- and 5-year-olds would be able to imagine the cause of the sound and choose the box made of the corresponding material, since such an ability would be a pre-requisite for their success in Experiment 1. Given that 3-year-olds and chimpanzees have been shown to infer the location of a reward based on the presence or absence of a sound cue in previous research (Call, 2004; Hill et al., 2012), we could predict that they would do so here, if they were able to match the sound made by the different materials to the appearance of the boxes. We therefore predicted that they would perform better than they did in Experiments 1 and 3.

Methods

Participants

Eleven chimpanzees (same as in Experiment 3) and a new group of 48 3–5-year-old children (16 in each age group) participated in this study (**Table 1**). Four additional children that were tested were excluded from the study due to parental interference (2) and refusal to complete the task (2).

Materials

The metal and wooden boxes from Experiment 3 were used. The boxes were covered with lids to block subjects' view. The ball was made from thermoplastic (1.30 cm in diameter). A barrier was used to occlude the hiding and shaking events.

Procedure

Children

The experimenter placed the boxes (approximately 15 cm apart from each other) and the ball on the table and introduced the task to the children: "In this game I have these two boxes. Now I will put a sticker in the ball and I will hide the ball in one of them. If you can find the ball, you will win the sticker!" Then the experimenter put the barrier in between and hid the ball in a box and shook it for approximately 5 s; and said "Here is a clue!" Children could see the arms of the experimenter but not the box being shaken. Then the experimenter placed the boxes in their original positions, removed the barrier and asked "Which box do you want to open?" Children received ten trials. The location of the boxes were counterbalanced.

At the end of the task, the experimenter asked children how they found the ball as in Experiment 1.

Chimpanzees

The procedure was the same for chimpanzees as in children apart from verbal instructions. Chimpanzees received 10 trials per session and testing continued until the subjects selected the correct side 16 out of 20 trials or more or until 10 sessions were completed.

Scoring and Analysis

A second examiner coded 20% of the videos for reliability, *Kappa* = 1.00, $p < 0.001$ for both children and chimpanzees. The full model based on the child data comprised of age, sex, and trial number as fixed effects. The full model of the chimpanzee data comprised of age, sex, session and trial numbers as fixed effects. For both models, ID and the side of the box were the random effects. We included random slopes of trial (and session numbers for chimpanzees) within ID but left out the correlation parameters between random intercepts and random slopes terms. The full and null model comparisons were done using a likelihood ratio test. In order to explore performance in this experiment against chance level ($p = 0.05$) we used one-sample t-tests.

Both models for children and chimpanzees were stable and there were no issues with regards to overdispersion (*dispersion parameter for children: 0.71; for chimpanzees: 0.99).* There was no multicollinearity issue for child data; however, age and sex predictors resulted in collinearity in chimpanzee data. Therefore, sex was dropped from the model.

Results

Children

The full model was significantly different from the null model [$\chi^2(4) = 16.62$, $p < 0.01$]. There was a significant effect of age, [$\chi^2(2) = 15.26$, $p < 0.001$], no effect of sex [$\chi^2(1) = 0.142$, $p = 0.706$] and no effect of trial number [$\chi^2(1) = 1.21$, $p = 0.271$] (**Supplementary Table S6**). The pairwise comparisons between age groups showed that there was a significant difference between the performances of 3- and 5-year-olds (GLMM, user-defined contrasts, z = 3.87, $p < 0.001$); no differences between 3- and 4-year-olds ($z = 1.89$, $p = 0.140$) nor between 4 and 5-year-olds ($z = 2.17$, $p = 0.08$) (see **Table 1** for means). Three-year-olds performed at chance [$M = 0.54$, 95% CI [0.44, 0.64], $t(15) = 0.94$, $p = 0.362$], 4- and 5-year-olds were significantly above chance; [$M = 0.69$, 95% CI [0.58, 0.81], $t(15) = 3.61$, $p < 0.01$] and [$M = 0.83$, 95% CI [0.71, 0.95], $t(15) = 5.91$, $p < 0.001$] respectively (**Figure 7**).

Table 4 summarizes the percentage of children in each age group based on their responses to the open-ended question. When the replies were lumped into three explanation categories as in Experiment 1, there was a significant relationship between age groups and explanations [$\chi^2(4) = 20.46$, $p < 0.001$]. The majority of the 3-year-olds were in the "no idea" category (88%) and only 1 (6%) gave the correct explanation. Among 4-year-olds, 40% were in the "no idea," 27% were in the "wrong idea" categories but 33% of them gave correct explanations. Among 5-year-olds only 31% were in the no or wrong idea categories and the majority (60%) were able to provide the correct explanation.

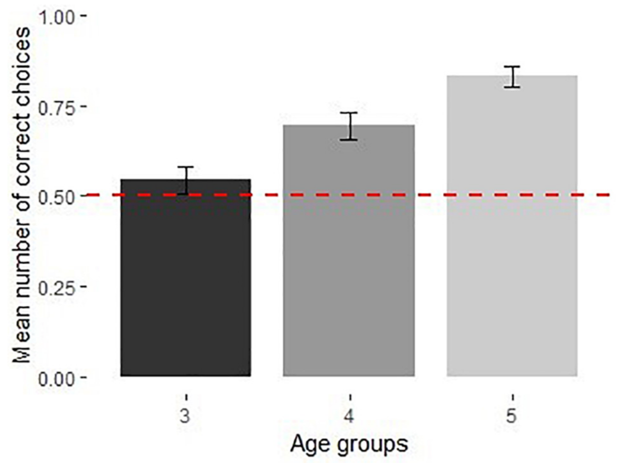


FIGURE 7 | Performance of 3–5-year-olds in shaken boxes in Experiment 4 (N = 48). Dotted line shows chance level performance (p = 0.05), error bars represent *SE*.

TABLE 4 | Percentage of children who gave the following explanations in response to the question "How did you know where the ball was?" in Experiment 4 (N = 48).

	Shaken boxes		
Explanations	**3 yo**	**4 yo**	**5 yo**
No explanation	88%	40%	13%
Wrong idea	6%	27%	19%
Correct explanation	6%	33%	69%

In order to examine whether children's reports matched with their performance, the performance of the 11 children who referred to different sounds/materials was compared with the performance of an age-matched group who gave other explanations. Those who referred to different sounds/materials performed significantly better ($M = 0.93$, $SE = 0.04$) than those who gave other explanations [$M = 0.62$, $SE = 0.06$), $t(20) = 4.26$, $p < 0.001$].

Chimpanzees

The full model for the chimpanzee data did not differ from the null model [$\chi^2(3) = 2.41$, $p = 0.492$] (**Supplementary Table S7**). Chimpanzees performed at chance level [$M = 0.50$, 95% CI [0.46, 0.53], $t(10) = -0.20$, $p = 0.844$] (**Figure 8**). However, one subject passed the shaken boxes condition in the 8th session. All but three individuals were side biased.

Discussion

Three-year-olds and chimpanzees could not infer the location of the ball based on auditory evidence about the material of a shaken box. In line with the previous findings from Experiment 1, we found that 4- and 5-year-olds performed significantly above chance level, corroborating the conclusion that by 4 years of age children are capable of reasoning about evidence to detect unseen causes in the absence of linguistic scaffolding.

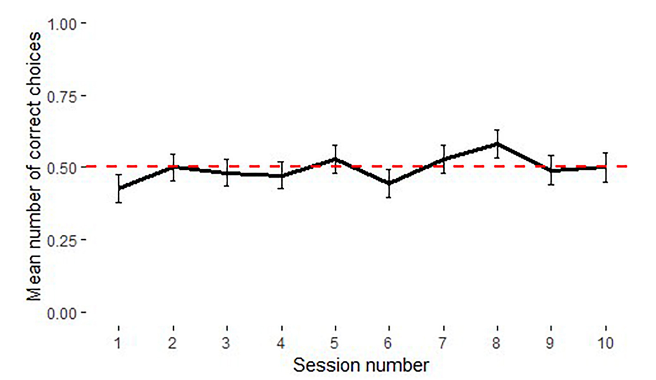


FIGURE 8 | Performance of chimpanzees in shaken boxes across sessions in Experiment 4 (*N* = 11). Dotted line shows chance level performance (*p* = 0.05), error bars represent *SE*.

In our last experiment we explored whether 3-year-olds' performance would improve with the addition of causal language as suggested by previous literature (Bonawitz et al., 2010; Butler and Markman, 2012; Lane and Shafto, 2017). We used the shaken boxes paradigm but this time provided cues to the causal structure of the task verbally.

EXPERIMENT 5: FOLLOW UP WITH CAUSAL LANGUAGE

Methods

Participants

A new group of 28 3-year-old children participated. There were equal numbers of boys and girls (**Table 1**). Three additional children that were tested were excluded from the study due to refusal to complete the task (2), and difficulties with language (1).

Materials

Same boxes were used as in Experiment 4.

Procedure

The procedure was the same as in Experiment 4 with the only exception of the question we asked children to locate the ball. Instead of "Which box do you want to open?" the experimenter asked "Which box did I shake?"

A second examiner categorized children's answers into five different types of explanations, *Kappa* = 0.97 [95% CI, 0.94, 0.99], $p < 0.001$.

Scoring and Analysis

A second examiner coded 20% of the videos for reliability, *Kappa* = 1.00, $p < 0.001$. In order to see the influence of a causal question on 3-year-olds' performance, we merged the data from Experiment 4 with the current data. Our model consisted of question type (non-causal as in Experiment 4/causal), trial number and sex as fixed effects; ID and the side of the boxes as random effects. We also included random slopes of trial number within ID, as well as the correlation parameters between random intercepts and random slopes terms (Schielzeth and Forstmeier, 2009; Barr et al., 2013). The full model was compared to a null model which included only the random effects using a likelihood ratio test.

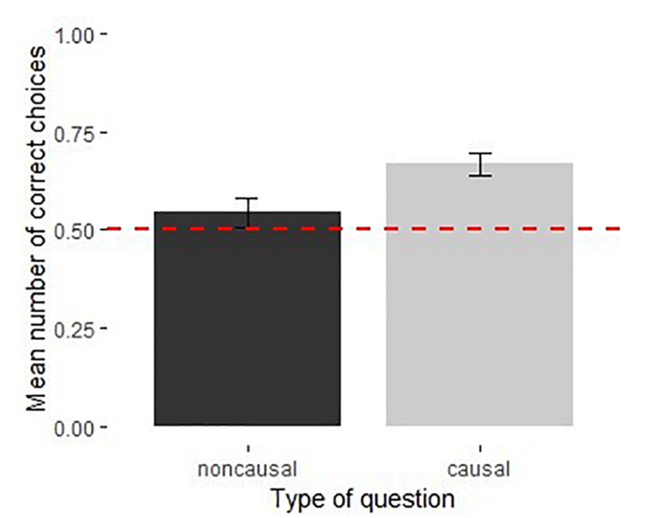


FIGURE 9 | Performance of 3-year-olds in shaken boxes when they were asked a non-causal (*N* = 16, Experiment 4) vs. a causal question (*N* = 28, Experiment 5). Dotted line shows chance level performance (*p* = 0.05), error bars represent *SE*.

The model was stable with regards to the predictors, there were no issues with regards to overdispersion (*dispersion parameter:* 0.85), nor multicollinearity.

We used one-sample *t*-test to examine whether children's performance was significantly different from chance level ($p = 0.05$).

Results

The full model was not significantly different from the null model [$\chi^2(3) = 5.44$, $p = 0.142$] (**Supplementary Table S8**). However, we found that 3-year-olds performed significantly above chance levels in the follow-up [$M = 0.67$, 95% CI [0.57, 0.76], $t(27) = 3.60$, $p < 0.01$] as opposed to their chance level performance in the absence of causal language (**Figure 9**).

The majority of the 3-year-olds were in the "no idea" category (71%) but 18% gave the correct explanation.

Discussion

Three-year-olds performed significantly above chance levels when they were asked a question that hinted at the causal structure of the shaken boxes task as opposed to chance level performance in Experiment 4. Even though the majority still could not explain how they found the ball, more children than in Experiment 1 gave the correct explanation. These findings showed that 3-year-olds were able to distinguish the auditory stimuli, and the peripheral demands of remembering what they heard and matching the sound with the box were not too high. However, this experiment does not explain how exactly verbal framing of the task facilitated performance. One possibility is that, the causal question boosted their performance through highlighting the problem; hence, scaffolding their ability to make inferences. Another possibility is that, by asking a question like "Which box did I shake?" we simplified the task such that it reduced the need for children to seek a causal explanation for the sounds they heard.

GENERAL DISCUSSION

We presented 3–6-year-old children and chimpanzees with a novel, natural causation task where they needed to use indirect evidence (auditory cues) to locate a reward in the absence of either a directly-perceivable causation relationship, or a verbal instruction to look for the cause of an outcome. By 4-years of age, children were able to make causal inferences based on evidence alone. Importantly, they only did so in the causally-plausible condition in which the falling ball could have caused the different sounds, rather than when the sound cues preceded the dropping of the ball and therefore bear an arbitrary relationship to its final location. In Experiment 2, we corroborated these findings by eliminating a simpler explanation for better performance in the causal condition than in the arbitrary one. Six-year-olds performed equally well and above chance levels in both causal and arbitrary conditions. This was in line with previous evidence showing that they are able to use arbitrary cues as meaningful symbols to solve a problem, in addition to detecting a causal structure from data. On the other hand, 3-year-olds and chimpanzees failed the task (Experiment 1 and 3). In Experiment 4, when the task was simplified to inferring the location of a reward in a metal and a wooden box based on the sounds it made, the performance of younger children and chimpanzees did not improve. But when the task was framed using causal language, 3-year-olds performed above chance levels. We discuss these findings and their implications in turn.

In the absence of causal instruction, 4- and 5-year-old children were able to use indirect auditory cues to locate a reward when there was a plausible causal structure to the task. They performed worse when the cues were arbitrarily related to the location of the reward. Children's explanations corroborated these findings: they referred to different sounds and materials in the causal condition more than the arbitrary condition and those who referred to different sounds outperformed their peers who gave other explanations.

Similar to the 4- and 5- year-olds, 6-year-olds passed the causal condition, but in contrast to the younger children they performed equally well in the arbitrary condition. This was as predicted based on similar findings with this age group in previous studies (DeLoache, 2004; Mayer et al., 2014). We suspect that older children solved the task because they interpreted the arbitrary cues as symbolic. DeLoache refers to this ability as holding dual representations: representing the symbol as an object/event by itself and also as a cue that stands for something else. In the arbitrary condition of our task, the metal and the wooden sounds that came before the ball was dropped had no causal relevance to the task, however, they could be treated as symbols that cued the child to which box the ball would be in since they were always predictive of the ball's location. Holding dual representations is cognitively challenging since one has to ignore the fact that it is causally irrelevant given the task but it may point to some information that is useful in order to solve the problem. For this reason DeLoache (2004) argued that the use of symbolic knowledge emerges fairly late in development, especially in the absence of verbal scaffolding.

Three-year-old children and chimpanzees did not discriminate between the conditions, did not pass either of them and were more likely to be side biased. This could reflect a "true negative": perhaps 3-year-olds and chimpanzees do not spontaneously make inferences about unseen causes when dealing with this kind of indirect evidence. Previous research has shown causal reasoning abilities in 3-year-olds in the context of direct causal relations and/or with explicit verbal scaffolding (Gopnik et al., 2001; Sobel et al., 2004; Bonawitz et al., 2010), but when they were presented with indirect causal structures such as a block activating a machine at a distance (Kushnir and Gopnik, 2007) or a task required them to represent prior knowledge to solve a problem (Sobel et al., 2004), 3-year-olds performed at chance level. Although suggestive of inferential reasoning abilities, the studies conducted with chimpanzees (Call, 2004; Hanus and Call, 2008) were criticized for not eliminating simpler associative explanations (Penn and Povinelli, 2007) or simplifying the task largely by using food itself as a cue (Völter and Call, 2014). We found no evidence to suggest that chimpanzees imagined or reasoned about the unseen causes involved in this study based on evidence alone. At face value, these findings may support a number of past claims in the literature (Penn and Povinelli, 2007): that non-human primates do not engage in inferential causal reasoning.

However, as with many negative findings, interpreting these results is not straightforward. One explanation for the failure of chimpanzees could be that the initial training we implemented were not sufficient to build the necessary knowledge for solving the problem. We used the transparent channels and sound-making boxes training separately to provide them with the required information for solving the test. One could argue that, integrating these two pieces of information might have been challenging. An alternative would be to incorporate these two together (i.e., tracking the movement of an object based on two different sounds in a forked apparatus). However, providing animals with training that is highly similar to the test phase makes it difficult to rule out associative explanations for success. In addition, chimpanzees' failure in Experiment 4 with shaken boxes despite the repeated experience with wood and metal sound-making boxes makes us more confident that the lack of prior experience was not the limiting factor.

Another explanation for the failure of chimpanzees could be limitations in executive function, specifically attention and working memory, which could mask or constrain their ability to reason. All of our tasks required subjects to integrate prior knowledge about object-object interactions with evidence to make inferences, and to keep track of transient auditory cues and match them with two different boxes. Although there is evidence that chimpanzees are sensitive to different sounds (Slocombe et al., 2009), capable of cross-modal matching (Davenport et al., 1973; Hashiya and Kojima, 2001), and inferring rewards based on auditory information (Call, 2004), performing all of these tasks at once may have overloaded their attention and working memory capacities. In support of this argument, Call (2007) found that; great apes failed to integrate information about the quality and the size of the reward when they were trying to locate a desirable piece of food under one of the two slanted boards although they were capable of doing these two tasks separately.

On the other hand, 3-year-olds' failure to succeed cannot easily be explained based on limitations in executive functions. It is true that our first experiment might have been challenging for young preschoolers, as it relied on integrating knowledge about channels and the sound of different materials that were not visible. The cognitive control abilities such as attention shifting and working memory undergo significant changes between the ages three and four (Frye et al., 1995; Zelazo and Muller, 2002) and this may influence 3-year-olds' performance when dealing with tasks where they need to keep track of multiple pieces of information. Previous research has shown how task difficulty may hamper performance of this age group (Hill et al., 2012). However, when we reduced the task difficulty largely with the removal of the channels and use of conspicuously metal and wooden boxes, it did not help 3-year-olds. On the other hand, they were able to solve the exact same task when causal language was involved, showing that the demands on executive functions were not too high. They were able to distinguish the two sounds, remember them at the time of choice and match them with the correct box. This brings us to the function of causal language.

How exactly verbal framing facilitates 3-year-olds' performance remains unclear. One possibility is that 3-year-olds fundamentally have the same cognitive machinery as 4- and 5-year-olds. Causal instructions/questions only highlight the problem among other irrelevant stimuli. In Experiment 5, as opposed to Experiment 4, children no longer needed to imagine that the boxes were shaken and this was causing the sounds they heard. This information was provided by the experimenter and hence they only needed to focus on the evidence to detect the true cause without imagining unseen actions or object-object interactions. Their true capacity was brought out by this verbal scaffold. The second possibility is that, with such causal questions the task is no longer measuring causal reasoning. From this point of view, the children were not required to make spontaneous inferences about evidence anymore but were asked to match a sound with the correct box. Which of these interpretations better explain the difference we find between Experiment 4 and 5 is an open and an interesting question that requires further research.

Overall, children's explanations about how they found the ball were in line with their problem-solving performance. First, there were more children in the causal condition than in the arbitrary one who said they found the ball based on different sounds it made in different boxes. Second, these children's explanations aligned with their performance: they performed better than those who referred to other explanations. Third, most of the 3-year-olds who performed at chance levels in both conditions either could not provide a verbal explanation or said they did not know how to find the ball implying that they found the task challenging. In addition when 3-year-olds' performance improved in the last experiment with causal language, this was reflected in their verbal explanations too: there were more children who gave correct explanations compared to Experiment 4. However, the majority still found it difficult, implying that linguistic expression is still developing. This association between explanation and problem-solving measures has been found in previous research. For example when children were prompted to explain what they observed (i.e., how a toy worked), they explored inconsistent outcomes and engaged in hypothesis testing strategies (Legare, 2012); and focused more on causal properties than on perceptual features of the evidence (Legare and Lombrozo, 2014; Walker et al., 2014). It has been argued that explaining promotes learning because it requires one to integrate evidence with prior beliefs (Lombrozo, 2006) and hence placing observations in the context of a larger and coherent framework (Wellman and Liu, 2007). Therefore, if some children were engaging in self-explanation while trying to solve this task, this could explain why they performed better than their peers. Further work could test this notion by prompting children to seek an explanation for the sounds, to see if this improves performance in 4- and 5- year-olds, which, while above chance, was not by any means at ceiling level.

To conclude, this work contributed to the developmental and comparative literature by introducing a novel paradigm that contrasts learning in a causal and an arbitrary context without the need for verbal instruction. We argue that our results are in line with previous suggestions that by 4-years of age, children are able to use evidence to detect unseen causes. It is possible that this stems from a tendency to seek causal explanations even in the absence of instruction to do so. Studies on exploratory play in young children conducted by Schulz and colleagues provide similar evidence that 4-year-old children are actively seeking out causal explanations. For instance, when provided with ambiguous information about how a toy worked, they spontaneously explored the toy's function rather than playing with a new toy as opposed to when the function was unambiguous (Schulz and Bonawitz, 2007); and they also conducted informative interventions (Cook et al., 2011). However, further work is needed to determine the reasons for the negative results found with younger children and chimpanzees. One possible avenue for future research will be the use of visual cues instead of auditory cues with a similar paradigm. This might improve performance of preschoolers and chimpanzees by lowering the cognitive load associated with tracking and remembering transient auditory cues.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of St Andrews Teaching and Research Ethics Committee. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. The animal study was reviewed and approved by University of St Andrews School of Psychology and Neuroscience Ethics Committee.

AUTHOR CONTRIBUTIONS

ZC performed the data collection, statistical analysis and wrote the first draft of the manuscript. All authors contributed to the design of the studies, revisions, read and approved the submitted version.

FUNDING

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 639072).

ACKNOWLEDGMENTS

We thank Robert MacKenzie for constructing the apparatus, Yanqiong Pan, Laras Yuniarto and Joao Pedro Sobrinho Ramos for reliability coding, Dundee Science Centre, Edinburgh Zoo and nurseries in St Andrews, Dundee and Cupar for letting us run our tests; and the animal caretakers of the Zoo Leipzig.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00872/full#supplementary-material

REFERENCES

Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computations of conditional probability statistics by 8-month-old infants. *Psychol. Sci.* 9, 321–324.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Barnet, R. C., Arnold, H. M., and Miller, R. R. (1991). Simulatenous conditioning demonstrated in second-order conditioning: evidence for similar associative structure in forward and simulataneous conditioning. *Learn. Motiv.* 22, 253–268. doi: 10.1016/0023-9690(91)90008-v

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., Machler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48.

Bonawitz, E. B., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., et al. (2010). Just do it? Investigating the gap between prediction and action in toddlers' causal inferences. *Cognition* 115, 104–117. doi: 10.1016/j.cognition.2009.12.001

Buchanan, D. W., and Sobel, D. M. (2011). Mechanism-based causal reasoning in young children. *Child Dev.* 82, 2053–2066. doi: 10.1111/j.1467-8624.2011.01646.x

Buehner, M. J., and May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Think. Reason.* 8, 269–295. doi: 10.1080/13546780244000060

Buehner, M. J., and May, J. (2003). Rethinking temporal contiguity and the judgement of causality: effects of prior knowledge, experience, and reinforcement procedure. *Q. J. Exp. Psychol.* 56A, 865–890. doi: 10.1080/02724980244000675

Bullock, M., and Gelman, R. (1979). Preschool children's assumptions about cause and effect: temporal ordering. *Child Dev.* 50, 89–96.

Bullock, M., Gelman, R., and Baillargeon, R. (1982). "The development of causal reasoning," in *The Developmental Psychology of Time*, ed. W. J. Friedman (New York, NY: Academic Press), 209–254.

Butler, L. P., and Markman, E. M. (2012). Finding the cause: verbal framing helps children extract causal evidence embedded in a complex scene. *J. Cogn. Dev.* 13, 38–66. doi: 10.1080/15248372.2011.567201

Call, J. (2004). Inferences about the location of food in the great apes (*Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla*, and *Pongo pygmaeus*). *J. Comp. Psychol.* 118, 232–241. doi: 10.1037/0735-7036.118.2.232

Call, J. (2007). Apes know that hidden objects can affect the orientation of other objects. *Cognition* 105, 1–25. doi: 10.1016/j.cognition.2006.08.004

Carey, S. E., and Spelke, E. (1996). Science and core knowledge. *Philos. Sci.* 63, 515–533.

Cook, C., Goodman, N., and Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition* 120, 341–349. doi: 10.1016/j.cognition.2011.03.003

Davenport, R. K., Rogers, C. M., and Russell, I. S. (1973). Cross-modal perception in apes. *Neuropsychologia* 11, 21–28. doi: 10.1016/0028-3932(73)90060-2

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends Cogn. Sci.* 8, 66–70. doi: 10.1016/j.tics.2003.12.004

Frye, D., Zelazo, P. D., and Palfai, T. (1995). Theory of mind and rule based reasoning. *Cogn. Dev.* 10, 483–527. doi: 10.1016/0885-2014(95)90024-1

Gelman, S. A. (2009). Learning from others: children's construction of concepts. *Annu. Rev. Psychol.* 60, 115–140. doi: 10.1146/annurev.psych.59.103006.093659

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* 111, 3–32. doi: 10.1037/0033-295x.111.1.3

Gopnik, A., Sobel, D. M., Schulz, L. E., and Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Dev. Psychol.* 37, 620–629. doi: 10.1037/0012-1649.37.5.620

Gopnik, A., and Wellman, H. M. (2012). Reconstructing constructivism: causal models, bayesian learning mechanisms, and the theory theory. *Psychol. Bull.* 138, 1085–1108. doi: 10.1037/a0028044

Hanus, D., and Call, J. (2008). Chimpanzees infer the location of a reward based on the effect of its weight. *Curr. Biol.* 18, 370–372.

Harris, P. L., and Koenig, M. A. (2006). Trust in testimony: how children learn about science and religion. *Child Dev.* 77, 505–524. doi: 10.1111/j.1467-8624.2006.00886.x

Hashiya, K., and Kojima, S. (2001). "Hearing and auditory–visual intermodal recognition in the chimpanzee," in *Primate Origins of Human Cognition and Behavior*, ed. T. Matsuzawa (Berlin: Springer-Verlag), 155–189. doi: 10.1007/978-4-431-09423-4_8

Hill, A., Collier-Baker, E., and Suddendorf, T. (2012). Inferential reasoning by exclusion in children (*Homo Sapiens*). *J. Comp. Psychol.* 126, 243–254. doi: 10.1037/a0024449

Kirkham, N. Z., Slemmer, J. A., and Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition* 83, 4–5. doi: 10.1016/S0010-1970277(02)00004-5

Köhler, W. (1925). *The Mentality of Apes*. London: Routledge & Kegan Paul.

Kushnir, T., and Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: preschoolers use new contingency evidence to overcome prior spatial assumptions. *Dev. Psychol.* 43, 186–196. doi: 10.1037/0012-1649.43.1.186

Lane, J. D., and Shafto, P. (2017). Young children's attributions of causal power to novel invisible entities. *J. Exp. Child Psychol.* 162, 268–281. doi: 10.1016/j.jecp.2017.05.015

Legare, C. H. (2012). Exploring explanation: explaining inconsistent evidence informs exploratory, hypothesis-testing behaviour in young children. *Child Dev.* 83, 173–185. doi: 10.1111/j.1467-8624.2011.01691.x

Legare, C. H., Gelman, S. A., and Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Dev.* 81, 929–944. doi: 10.1111/j.1467-8624.2010.01443.x

Legare, C. H., and Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *J. Exp. Child Psychol.* 126, 198–212. doi: 10.1016/j.jecp.2014.03.001

Legare, C. H., Wellman, H. M., and Gelman, S. A. (2009). Evidence for an explanation advantage in naïve biological reasoning. *Cogn. Psychol.* 58, 177–194. doi: 10.1016/j.cogpsych.2008.06.002

Limongelli, L., Boysen, S. T., and Visalberghi, E. (1995). Comprehension of cause-effect relations in a tool-using task by chimpanzees (*Pan troglodytes*). *J. Comp. Psychol.* 109, 18–26. doi: 10.1037/0735-7036.109.1.18

Lombrozo, T. (2006). The structure and function of explanations. *Trends Cogn. Sci.* 10, 464–470. doi: 10.1016/j.tics.2006.08.004

Mayer, C., Call, J., Albiach-Serrano, A., Visalberghi, E., Sabbatini, G., and Seed, A. (2014). Abstract knowledge in the broken-string problem: evidence from nonhuman primates and pre-schoolers. *PLoS One* 9:e108597. doi: 10.1371/journal.pone.0108597

Michotte, A. (1963). *The Perception of Causality*. Oxford: Basic Books.

Miller, R. R., and Barnet, R. C. (1993). The role of time in elementary associations. *Curr. Dir. Psychol. Sci.* 2, 106–111. doi: 10.1111/1467-8721.ep10772577

Penn, D. C., Holyoak, K. J., and Povinelli, D. J. (2008). Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* 31, 109–178.

Penn, D. C., and Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind". *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 362, 731–744. doi: 10.1098/rstb.2006.2023

Povinelli, D. J. (2000). *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works*. Oxford: Oxford University Press.

Rankin, M. L., and McCormack, T. (2013). The temporal priority principle: at what age does this develop? *Front. Psychol.* 4:178. doi: 10.3389/fpsyg.2013.00178

Reber, A. S. (1989). Implicit learning and tacit knowledge. *J. Exp. Psychol. Gen.* 118, 219–235.

Schielzeth, H., and Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behav. Ecol.* 20, 416–420. doi: 10.1093/beheco/arn145

Schulz, L. E., and Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Dev. Psychol.* 43, 1045–1050. doi: 10.1037/0012-1649.43.4.1045

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., and Jenkins, A. C. (2008). Going beyond the evidence: abstract laws and preschoolers' responses to anomalous data. *Cognition* 109, 211–223. doi: 10.1016/j.cognition.2008.07.017

Schulz, L. E., and Sommerville, J. (2006). God does not play dice: causal determinism and preschoolers' causal inferences. *Child Dev.* 77, 427–442. doi: 10.1111/j.1467-8624.2006.00880.x

Seed, A., Hanus, D., and Call, J. (2011). "Causal knowledge in corvids, primates and children: more than meets the eye?," in *Tool Use and Causal Cognition*, eds T. Mccormack, C. Hoerl, and S. Butterfill (Oxford: Oxford University Press), 89–111.

Seed, A. M., and Call, J. (2014). Space or physics? children use physical reasoning to solve the trap problem from 2.5 years of age. *Dev. Psychol.* 50, 1951–1962. doi: 10.1037/a0036695

Seed, A. M., and Mayer, C. (2017). "Problem solving," in *APA Handbook of Comparative Psychology*, Vol. 2, eds J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, and T. R. Zentall (Washington, DC: American Psychological Association), 601–625.

Shanks, D. R., and Dickinson, A. (1987). "Associative accounts of causality judgement," in *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 21, ed. G. H. Bower (Cambridge, MA: Academic Press), 229–261. doi: 10.1016/s0079-7421(08)60030-4

Shanks, D. R., Pearson, S. M., and Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *Q. J. Exp. Psychol. Section B* 41, 139–159.

Shultz, T. R. (1982). Rules of causal attribution. *Soc. Res. Child Dev.* 47, 1–51.

Siegel, M. H., Magid, R., Tenenbaum, J. B., and Schulz, L. E. (2014). "Black boxes: hypothesis testing via indirect perceptual evidence," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, Cambridge, MA.

Slocombe, K. E., Townsend, S. W., and Zuberbuhler, K. (2009). Wild chimpanzees (*Pan troglodytes schweinfurthii*) distinguish between different scream types: evidence from a playback study. *Anim. Cogn.* 12, 441–449. doi: 10.1007/s10071-008-0204-x

Sobel, D. M., and Sommerville, J. A. (2009). Rationales in children's causal learning from others' actions. *Cogn. Dev.* 24, 70–79. doi: 10.1016/j.cogdev.2008.08.003

Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2004). Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. *Cogn. Sci.* 28, 303–333. doi: 10.1016/j.cogsci.2003.11.001

Suddendorf, T., Crimston, J., and Redshaw, J. (2017). Preparatory responses to socially determined, mutually exclusive possibilities in chimpanzees and children. *Biol. Lett.* 13:20170170. doi: 10.1098/rsbl.2017.0170

Völter, C. J., and Call, J. (2014). Great apes (*Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo abelii*) follow visual trails to locate hidden food. *J. Comp. Psychol.* 128, 199–208. doi: 10.1037/a0035434

Völter, C. J., and Call, J. (2017). "Causal and inferential reasoning in animals," in *APA Handbooks in Psychology. APA Handbook of Comparative Psychology: Perception, Learning, and Cognition*, eds J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, and T. Zentall (Washington, DC: American Psychological Association), 643–671. doi: 10.1037/0000012-029

Walker, C. M., Lombrozo, T., Legare, C. H., and Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition* 133, 343–357. doi: 10.1016/j.cognition.2014.07.008

Wellman, H. M., and Liu, D. (2007). "Causal reasoning as informed by the early development of explanations," in *Causal Learning: Psychology, Philosophy and Computation*, eds A. Gopnik and L. Schulz (Oxford: Oxford University Press).

Zelazo, P. D., and Muller, U. (2002). "Executive function in typical and atypical development," in *Handbook of Childhood Cognitive Development*, ed. U. Goswami (Oxford: Blackwell), 445–469. doi: 10.1002/9780470996652.ch20

Conflict of Interest: 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 May 2020
doi: 10.3389/fpsyg.2020.01069

Causal Responsibility and Robust Causation

Guy Grinfeld[1], David Lagnado[2], Tobias Gerstenberg[3], James F. Woodward[4] and Marius Usher[1,5]*

[1] School of Psychology, Tel Aviv University, Tel Aviv, Israel, [2] Cognitive, Perceptual, and Brain Sciences Department, Experimental Psychology, University College London, London, United Kingdom, [3] Stanford University, Stanford, CA, United States, [4] Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA, United States, [5] Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel

OPEN ACCESS

Edited by:
York Hagmayer,
University of Göttingen, Germany

Reviewed by:
Nadya Vasilyeva,
Princeton University, United States
Momme Von Sydow,
University of Siegen, Germany

***Correspondence:**
David Lagnado
d.lagnado@ucl.ac.uk

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

Received: 20 August 2019
Accepted: 27 April 2020
Published: 27 May 2020

Citation:
Grinfeld G, Lagnado D,
Gerstenberg T, Woodward JF and
Usher M (2020) Causal Responsibility
and Robust Causation.
Front. Psychol. 11:1069.
doi: 10.3389/fpsyg.2020.01069

How do people judge the degree of causal responsibility that an agent has for the outcomes of her actions? We show that a relatively unexplored factor – the robustness (or stability) of the causal chain linking the agent's action and the outcome – influences judgments of causal responsibility of the agent. In three experiments, we vary robustness by manipulating the number of background circumstances under which the action causes the effect, and find that causal responsibility judgments increase with robustness. In the first experiment, the robustness manipulation also raises the probability of the effect given the action. Experiments 2 and 3 control for probability-raising, and show that robustness still affects judgments of causal responsibility. In particular, Experiment 3 introduces an Ellsberg type of scenario to manipulate robustness, while keeping the conditional probability and the skill deployed in the action fixed. Experiment 4, replicates the results of Experiment 3, while contrasting between judgments of causal strength and of causal responsibility. The results show that in all cases, the perceived degree of responsibility (but not of causal strength) increases with the robustness of the action-outcome causal chain.

Keywords: causality and responsibility, attributions of responsibility, robust causation, causal contingency and stability, epistemic perspective

INTRODUCTION

The causal responsibility an agent has for the effects of her actions is thought to play a major role in the attribution of the agent's legal, moral and even criminal responsibility (Hart and Honoré, 1959; Tadros, 2005; Moore, 2009; Lagnado and Gerstenberg, 2017; Usher, 2018). Indeed, causal responsibility is a necessary condition for the ascription of legal responsibility (Hart and Honoré, 1959; Tadros, 2005). Research in moral psychology has identified general cognitive processes such as causal and intentional attributions to explain patterns of responsibility judgments in both moral and non-moral domains (Cushman and Young, 2011; see also Spranca et al., 1991; Royzman and Baron, 2002; Mikahil, 2007; Waldmann and Dieterich, 2007; Lagnado and Channon, 2008; Baron and Ritov, 2009; Greene et al., 2009). For example, Cushman and Young (2011) show that action versus omission and means versus side-effect differences in moral judgments are mediated by their effects on non-moral representations of causal and intentional attributions.[1] Similarly, Mikahil (2007) accounts for the means versus side-effect distinction in terms of action plans which specify generic rather than morally specific reasoning (see also Kleiman-Weiner et al., 2015). Furthermore, the reduced judgments of responsibility people typically make for agents who were forced to act or were manipulated by others (Sripada, 2012; Phillips and Shaw, 2015) are best accounted for by the manipulation "bypassing" the agent's mental states (Murray and Lombrozo, 2017). To explain these results, Murray and Lombrozo (2017) relied on the notion of *counterfactual robustness*, which is central to our investigation and on which we elaborate below.

While causation is basic in all sciences, its proper understanding has been the subject of intensive recent development and debate in philosophy (Pearl, 2000; Woodward, 2003; Hitchcock, 2007; Halpern, 2016). A key distinction is between type and token level causation. The former applies to laws and generalizations (such as smoking causes cancer), while the latter applies to particular cases (John's smoking caused his cancer). The concept of causal responsibility is typically grounded in the notion of token (or actual) causation. This means that to know that an agent is causally responsible for an effect, we need to know that she actually caused it (in that particular case). However, while the classical analysis of actual causation is based on a *necessity* counterfactual – if C had not occurred, then E would not have occurred (Hume, 1748; Lewis, 1973) – which is an all-or-none concept (the counterfactual is either true or false) – judgments of causal and legal responsibility are *graded*. For example, an agent generally bears more responsibility for an effect that is a direct outcome of her action, than for an effect that results at the end of a long causal chain (Brickman et al., 1975; Spellman, 1997; McClure et al., 2007; Lagnado and Channon, 2008; Hilton et al., 2010). This is aptly illustrated in the Regina v. Faulkner (1877) legal case, in which a lit match aboard a ship caused a cask of rum to ignite, causing the ship to burn and resulting in a large financial loss by Lloyd's insurance, leading to the suicide of a (financially ruined) insurance executive.[2]

Experimental studies of actual causation have also shown that judgments of causal strength and of causal responsibility vary with the typicality of the cause and the background conditions (Hilton and Slugoski, 1986; Kominsky et al., 2015; Samland and Waldmann, 2016; Icard et al., 2017; Gerstenberg et al., 2018) and also with the degree of causal redundancy (Gerstenberg and Lagnado, 2010; Zultan et al., 2012; Lagnado et al., 2013; Gerstenberg et al., 2015; Koskuba et al., 2018). For example, Kominsky et al. (2015) show that people are more likely to endorse an event (Alex's coin-flip coming up heads) as causally responsible for another event (Alex wins the game), when the contingency between the two is high (Alex wins if both the coin comes up heads and the sum of two dice being thrown is *greater than 2*) than when the contingency between the two events is low (Alex wins if both the coin comes up heads and the sum of two dice being thrown is *greater than* 11). Also, when there is causal redundancy, for example, when the action of several agents overdetermines an outcome (e.g., two marksmen are shooting a person; Lagnado et al., 2013), the responsibility of each agent is reduced the more overdetermined the outcome is.

The interventionist framework (Pearl, 2000; Woodward, 2003) provides a general framework for understanding causal claims at both type and token level (we focus on the latter here). On this approach, X causes Y if some potential manipulation of X would lead to a change in Y, under suitable background conditions (for full details see Woodward, 2003). Theoretical work within this framework has highlighted two ways to extend the classical analysis of causality to provide room for degrees of causal responsibility. The first involves a refinement of the *necessity* condition (Chockler and Halpern, 2004; Halpern and Hitchcock, 2015), which suggests that necessity is to be tested not only on counterfactuals that negate the cause and keep all other co-factors constant, but also on counterfactuals that can vary some of these co-factors. For example, in situations of redundant causation (e.g., two marksmen shooting a person), neither of the agents' actions is necessary for the outcome (either shot on its own was sufficient to kill the victim, so the victim would still have died, even if one of the marksmen hadn't shot). However, if we allow for a more flexible type of counterfactual test, where one marksman's action is assessed under the contingency where the other marksman does not shoot, then both marksmen can be counted as causes of the victim's death. Moreover, this extended counterfactual account fits with empirical data on graded causal judgments (Zultan et al., 2012; Lagnado et al., 2013).

The second way to introduce gradations of causal judgments involves a complementary causal condition: *robust sufficiency* (Lewis, 1973; Pearl, 1999; Woodward, 2006; Lombrozo, 2010; Hitchcock, 2012; Kominsky et al., 2015; Icard et al., 2017; Usher, 2018; Vasilyeva et al., 2018). Focusing on the simplest case, where we have one putative cause X of an effect Y, and a set of background circumstances B: X is robustly sufficient for Y if, given that X occurs, Y would still occur, even under various changes to the background circumstances. In contrast, the sufficiency of X for Y is non-robust (or highly *sensitive*) if, given X, Y would only occur under a very specific (narrow) set of background circumstances. Thus, we have a spectrum of degrees of robustness according to the range of background circumstances under which X would remain sufficient (i.e., be pivotal) for Y. This notion of robustness can be generalized to more complicated situations involving multiple causal factors (see Gerstenberg et al., 2015; Kominsky et al., 2015), and also to non-deterministic contexts where causes merely raise the probability of their effects (Hitchcock, 2017). Applied to the case of Regina vs. Faulkner, the causal chain from the lit match to the suicide of the insurance executive seems non-robust (and highly sensitive): it held only under this very specific set of background circumstances, and would have failed if only one of these factors had been different (see also Halpern and Hitchcock, 2015). Robust actions, on the other hand, are thought

[1] These two causal responsibility patterns are as follows: (i) Harm brought about by an action is deemed morally worse than harm brought about by an omission. (ii) People judge harm used as the necessary means to a goal to be worse than harm produced as the foreseen side-effect of a goal (Cushman and Young, 2011).

[2] The executive's widow sued for compensation, but it was ruled that the negligence of lighting the match was not a cause of his death.

to involve some degree of stability (low sensitivity) on the impact of background circumstances that are external to the action itself (Woodward, 2006).

Robust Sufficiency Versus Probability Raising

An alternative framework for quantifying causal strength employs the notion of *probability-raising* (Suppes, 1970; Cheng and Novick, 1992; Spellman, 1997; Fitelson and Hitchcock, 2011), where the strength of the relation between cause and effect corresponds to the degree to which the cause raises the probability of the effect (holding all else equal). This account usually focuses on type-level causal relations (Cheng, 1997), but has been extended to actual causation (Spellman, 1997; Cheng and Novick, 2005; Stephan and Waldmann, 2018). Although the framework suffers from notorious difficulties in distinguishing correlation from causation (Cartwright, 1989; Woodward, 2003; Pearl, 2009), it can be revamped to give a potential measure of causal strength (e.g., probability-raising through intervention).

While robust causes will often raise the probability of their effects more than non-robust causes, robustness and probability-raising are potentially distinguishable. For example, suppose that a doctor considers two possible drugs to treat a difficult medical condition that has a 20% chance of recovery if untreated. Drug X has a success rate of 60% in two possible background conditions (B1 and B2, whose presence is difficult to establish), while drug-Y has a 100% recovery in B1, but only 20% in B2. Overall, if we assume that B1 and B2 are equally probable, then both drugs yield the same recovery rate of (60% + 60%)/2 = (80% + 20%)/2 = 60%.[3] However, Drug X is more robust, because the relation between X and recovery holds under a greater number of background conditions, (B1 and B2 for X versus only B1 for Y). We will exploit this kind of example to de-confound robustness and probability-raising in our two latest experiments (3 and 4) that focus on the causal-responsibility of an agent for the effects of her actions, using a design similar to one recently employed by Vasilyeva et al. (2018), in the context of causal generalizations and causal explanations. Note that defining robust-sufficiency in terms of the number of background circumstances, rather than in terms of probability raising, has two important advantages. First, an agent may know of different background circumstances which moderate the relationship between the cause and the effect, but not have information about the probabilities with which those background circumstances occur (or about the probabilities of the effect conditional on the cause in those circumstances) and, hence may not have the information to make reliable judgments of probability raising. In such cases, the agent may still be guided by an estimate of the number of different circumstances in which the cause leads to the effect — that is, the robustness of the cause/effect relationship. Second, there are theoretical reasons to prefer causal relations that are invariant in various ways, in particular invariant to changes in background conditions (Cheng, 1997; Woodward, 2003).[4] We will return to the distinction between robustness and probability raising in the Discussion.

Judgments of Causation vs. Judgments of Responsibility

Our interest here is in judgments of causal responsibility that agents have for the outcomes of their actions, which are an essential component of the type of responsibility that is involved in judgments of praise and blame. While it is beyond the scope of our paper to offer (or test) a full theory of praise/blame (Malle et al., 2014; Halpern and Kleiman-Weiner, 2018), we note that a common assumption is that there are two components in credit/blame attribution: (i) an intentional one (the agent needs to have intended and foreseen the outcome of her action and the intention must not be the outcome of manipulation by another agent; Lagnado and Channon, 2008; Sripada, 2012; Phillips and Shaw, 2015), and (ii) a causal one: the agent must be causally responsible for the outcome in virtue of an action she did (or failed to do). In this study, we will only focus on the second component, by keeping the intention present and fixed. In that sense, our judgment of interest, *causal-responsibility*, is similar to judgments of *causal strength* (but note that we focus on judging the responsibility of agents for an outcome of their actions). Probing causal strength typically asks "to what extent X caused Y," while probing causal responsibility asks "how responsible is X for Y" (Sarin et al., 2017). Indeed, most studies in the field have probed participants with either of these measures with parallel effects (Murray and Lombrozo, 2017; Sarin et al., 2017). Moreover, Sytsma et al. (2012) have proposed that "the ordinary concept of causation, at least as applied to agents, is an inherently normative concept: Causal attributions are typically used to indicate something more akin to who is responsible for a given outcome than who caused the outcome in the descriptive sense of the term used by philosophers" (p. 815).

There are reasons, however, to expect that under certain conditions, causal and responsibility judgments may diverge, even when the agent's intention is held constant (Chockler and Halpern, 2004). Consider, for example, a doctor that administers drug X (60% recovery in both B1 and B2; robust to background conditions) or Y (100% recovery in B1 and 20% in B2; sensitive to background conditions), in the example above, and assume that the background circumstance B1 takes place (and B2 does not). In such a case, we believe that there is a good reason to expect a dissociation. While the causal strength between the doctor's action and the patient's recovery is likely to be higher in the case of drug Y (which in the actual circumstance, corresponding to B1, increases the probability of the effect by 80%, compared with an increase of only 40% for drug X), the causal responsibility

[3] As the baseline recovery probability (without treatment) is fixed, the two treatments also have the same impact in terms of ΔP and causal-power (Cheng, 1997).

[4] As discussed in Woodward (2003), there are a variety of invariance conditions, which a causal claim might satisfy. Robustness in the sense of invariance under changes in background conditions is just one such requirement. Other invariance requirements may be found, for example, in the characterization of causal power (Cheng, 1997), which rests on the assumption that causes will (at least in simple cases) conform to two invariance requirements — the power of the cause to produce an effect should be invariant under changes in the frequency with which the cause occurs and also invariant under changes in the frequency with which other causes of the effect occur.

attributed to the agent for the same event, is likely to be higher for substance X, as X is more robust, and thus the outcome is less sensitive to external circumstances. This distinction is consistent with a recent theory of actual causation proposed by Chockler and Halpern (2004), who distinguish between judgments of causality, responsibility and blame. For example, they consider the case of a person being shot by a firing squad of 10 marksmen, only one of whom has a live bullet. According to Chockler and Halpern, while only the marksman with the live bullet caused the death of the person, the *blame* is divided between all 10 marksmen, reflecting the *epistemic* uncertainty of the marksmen agents. Here, we propose that judgments of causal responsibility elicit the same epistemic perspective change as those of blame. While this differs from the distinction made by Chockler and Halpern, we believe that in the presence of the intentional component[5], the distinction between attributions of blame/praise and causal responsibility is more subtle and thus is beyond the scope of the present work. For that reason, we group these two aspects (causal-responsibility and credit/blame) together in this study.

There are two recent studies that are relevant, in particular, to our present work. The former examined judgments of credit and blame, which were shown to depend on two factors: (i) the degree of causal redundancy (or *pivotality*) and (ii) the amount of skill that people infer the agent to possess (Gerstenberg et al., 2018). People attribute more credit to an agent when their action has a higher contingency to the outcome, indicative of skill (rather than luck), and when their action is pivotal for bringing about the outcome. In another recent study, Vasilyeva et al. (2018) reported an effect of stability on judgments of causal generalizations and causal explanations for causal contingencies. In their study, participants made judgments under conditions of uncertainty about the factors that determine the causal contingency. Here, we aim to contrast judgments of causal responsibility and of causal strength for agents that bring about an effect in a robust vs. non-robust manner. Moreover, we will focus on situations in which the agents, but not the participants, face epistemic uncertainty. This allows us to test the impact of robustness under situations with minimum epistemic uncertainty (from the perspective of the judging participant), as well as tease apart robustness, causal strength, and causal responsibility (of agents for outcomes of their actions). We defer a more detailed discussion of the differences between our study and that of Vasilyeva et al. (2018) to the General Discussion.

Overview of the Paper

This paper aims to empirically test the impact of robustness on judgments of causal responsibility in human agency. To do so, in all our experiments, we probe the extent to which a human agent is judged to be causally responsible for the intended outcome that resulted from her action. In Experiment 1 (which probes both judgments of causal responsibility of causal strength), we manipulate robustness in terms of the number of background conditions under which the action would bring about the effect[6], and assess judgments of both causal responsibility and causal strength. However, as noted, robustness is often correlated with probability-raising, and Experiment 1 does not discriminate between these two factors. Thus, Experiments 2–4 probe causal responsibility and manipulate robustness while holding probability-raising constant. Experiment 3 teases apart the effects of robustness and skill on causal responsibility (Gerstenberg et al., 2018), by holding the agents' skill level constant, while varying the robustness of their actions. Finally, in Experiment 4, we replicate Experiment 3 and extend the results to cases of failure, contrasting judgments of causal responsibility and of causal strength.

EXPERIMENT 1

In this experiment, we probe robustness by manipulating the number of background conditions under which an action is likely to result in a positive outcome. To do so, we designed a game in which an agent throws a dart at a target to determine how many dice will be rolled (one, two, or three) by a computer, with the player winning the game if the sum total on the dice is six or greater. The number of dice rolled (one, two, or three) depends on the agent's dart throw in the following way: if the dart lands in one of the two inner circles (rings 5–6) three dice are rolled, if it lands in one of the two middle circles (3–4), two dice are rolled, and finally, if it lands in one of the two outer circles (1–2), one die is rolled. While under all three contingencies a win is possible, the degree of robustness increases with the number of dice rolled.[7] For the single die roll, there is only one configuration (a six) resulting in a win. When two dice are rolled, there are several more configurations that result in a win (e.g., 6,1; 5,1; 4,2; etc.; 26 out of 36 possible outcomes), and for three dice there are even more possible configurations (e.g., 2,2,2; 1,4,5; etc.; 206 out of 216 possible outcomes). In other words, the action (the dart throw) can cause the desired outcome (winning) under more, or less, background conditions (dice rolls). Observers in the experiment watched an animation in which an agent first throws a dart which lands on one of the rings (1–6) of the dartboard. After that, the corresponding number of dice were "randomly" rolled, the outcome of the roll revealed, and winning or losing declared. Half of the participants were asked to evaluate the degree of responsibility of the agent for the win/loss. The other half was asked to evaluate the causal strength between the agent's throw and the win/loss. In this experiment, since robustness and probability raising co-vary, while the agent's

[5] Variations in this component can explain important differences between blame and causal responsibility. For example, unlike an adult, a child who pokes at a gun's trigger out of curiosity will not be held culpable for resulting injury or death, given reasonable assumptions about the lack of relevant intentions (Shafer, 2000; Lagnado and Channon, 2008).

[6] There are many contexts in which it is unclear how to count the "number of background conditions" (see Phillips and Cushman, 2017). In such contexts, the notion of robustness is not uniquely defined. However, in Experiment 1, there is a very natural way of counting the number of relevant background conditions. The number of background conditions is also well-defined when one set of background circumstances is a proper subset of another.

[7] According to Woodward's conceptualization, robustness depends on the number of background conditions (BC) under which the desired outcome takes place, given the action (Woodward, 2006). In this case, the relevant BCs are the possible outcomes of the dice that equal or exceed 6.

epistemic perspective is held constant, we expect attributions of causal responsibility and causal strength to show similar patterns (cf. Vasilyeva et al., 2018). In Experiment 4, we identify a context in which the two measures dissociate.

We anticipated three possible patterns of results: (i) judgments of responsibility/causal strength will not be affected by the dart-ring outcome, implying that they are only sensitive to the outcome (success or failure), (ii) judgments will increase monotonically depending on which ring (1–6) the dart landed in, implying that they are mainly determined by the agent's perceived skill, (iii) judgments will increase in two steps, from ring 2 to 3, and from ring 4 to 5, implying that they are determined by the robustness of the dart throw (action)/game-win (effect) contingency to background conditions (outcomes of the dice rolls).

Method

Participants

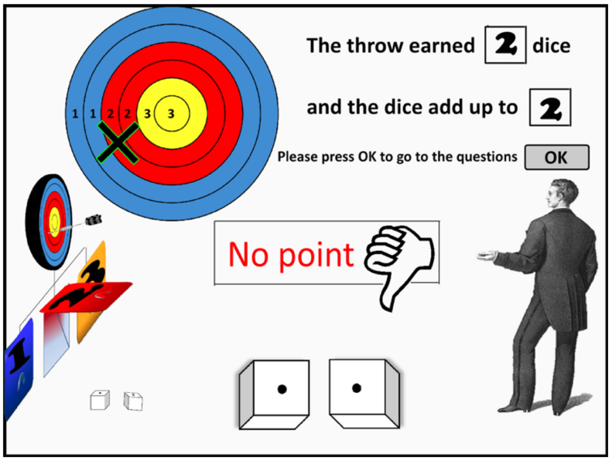
One hundred and two participants (46 females, 56 males; mean age = 33.7, *SD* = 9.7) took part in the experiment. Fifty of whom rated how much the agent's throw was a cause for winning or losing. The other 52 participants rated the agent's responsibility for the result. All participants were recruited via Amazon Mechanical Turk and received 1$ for participating.

Materials

Twenty video clips were created (see **Figure 1** for illustration), which varied in terms of the ring in which the dart lands (1–6), the associated number of dice rolled (1–3), and the total score of the dice (2–12). The profiles of each of the 20 clips are shown in **Table 1**. Note that a total score of 6 or larger corresponds to a win, and scores lower than 6 to a loss. Here, we focus on the responsibility for success, but we included cases of failure in order to balance success and failures to make the game credible. We also constructed the set so that the total score of 6 (the critical value for success) appears for all possible ring numbers and number of dice thrown. This allows for a straightforward comparison of the influence of these variables on people's ratings. When more than one dice was thrown, we included in addition winning cases in which the total scores were 9 or 12 (# of dice = 2, and # of dice = 3, respectively).

Procedure

After having received instructions, each participant watched the 20 clips in randomized order. For each clip, participants watched a video showing a player throwing a dart toward the board, and landing on one of the six rings (1–6). Depending on the result of the dart throw, a number of dice (one, two or three) were rolled, and the total on the dice was displayed, along with notification of a win or a loss of a point (see **Figure 1**). Participants were then asked to enter their evaluation of responsibility ("to what extent is the player responsible for wining/losing of this point") or of causal strength ("to what extent did the player cause the win/loss of this point"), by using a slider on a rating scale with endpoints labeled from '0 = not at all responsible (not at all the cause)' to '10 = completely responsible (completely the cause).'



The throw earned 2 dice

and the dice add up to 2

Please press OK to go to the questions

No point

FIGURE 1 | A screen-shot from the animation that participants watched in Experiment 1. The agent (on the right) throws the dart toward the board. The board is colored to illustrate the number of dice rolls earned. In this case, the arrow landed on Ring 3 (counting from the outside in), earning two dice. Each die landed on a 1, thus no point was won (a sum total for the dice of at least 6 was required to score a point).

Analysis

Our main focus is on the win data, and in particular, when the total dice-score was equal to six, as this was the only winning outcome in all dice conditions. We thus carried out an ANOVA on the impact of the number of dice thrown (1, 2, or 3), of the ring number (odd vs. even), and of the type of judgment (responsibility vs. causal strength). Note that the odd/even categorical variable of ring number, contrasts between ring values (1, 3, 5) vs. (2, 4, 6), and thus tested if ring value matters once we control for the number of dice thrown. For completeness, we also ran a linear mixed-effects model on all success trials (where the total score is equal to or greater than 6), in which we predicted the causation and responsibility ratings from three variables: number of dice thrown, ring number (odd vs. even), and total score, as well as a participant dependent intercept. Finally, we ran exactly the same analyses for the cases of failures. Here, we replaced cases in which total dice score was 6 with those in which it was 5 in the ANOVA, and we also a linear mixed-effects model all failure trials with number of dice thrown, ring number and total score as predictors.

Results

We start with the causal strength and causal responsibility judgments that are ascribed to the dart-throw agent for wins (total score $\geq$ 6). To illustrate the results, we plot in **Figure 2** the mean ratings (for the total-score value of 6, which is the only win-condition, that appears in combination with all ring-values) for judgments of causal strength (left panel), and for responsibility (right panel), as a function of number of dice and ring (odd/even).

To test the impact of number of dice and ring-parity, we carried out a 3 × 2 × 2 ANOVA with number of dice rolled (one/two/three) and the ring-parity (odd vs. even) as within-subject factors, and judgment type (causation vs. responsibility)

TABLE 1 | Profiles of the 20 clips used in Experiment 1.

Clip	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Ring	6	6	6	5	5	5	4	4	4	4	3	3	3	3	2	2	2	1	1	1
N. dice	3	3	3	3	3	3	2	2	2	2	2	2	2	2	1	1	1	1	1	1
Score	5	6	12	6	12	5	2	6	9	5	9	5	2	6	5	2	6	2	6	5
Result	L	W	W	W	W	L	L	W	W	L	W	L	L	W	L	L	W	L	W	L

Ring = ring that the dart landed in (1 = most outer ring, 6 = center ring), dice = number of dice rolled, score = what sum the dice added up to, result = loss (l) or win (w). Scores of 6 or higher resulted in a win.

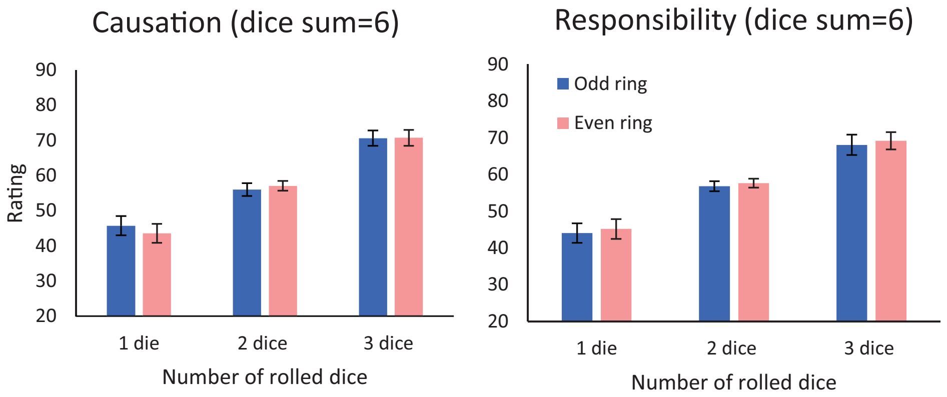


FIGURE 2 | Evaluations of causal strength (left) and of responsibility (right) as a function of the number of dice rolled and of the ring the dart hit: odd (blue) vs. even (pink). Error bars indicate a within-subject ±1 standard error of the mean (Cousineau, 2005).

as a between-subject factor. The results showed a significant main effect for the number of dice rolled, $F(2,200) = 55.21$, $p < 0.001$, $\eta_p^2 = 0.356$. *Post hoc* Bonferroni tests show that ratings for hitting the inner rings (three dice) were significantly higher than for the middle rings [two dice; $t(100) = 7.57$, $p < 0.001$, $d = 0.750$], which were significantly higher than outer rings [one dice; $t(100) = 6.01$, $p < 0.001$, $d = 0.596$]. No other main effects or interactions were significant (see **Supplementary A** for a report of the non-significant results).

This result was confirmed by a linear mixed effects model with random intercepts for each participant on all success outcomes trials, which showed that only the number of dice thrown was a significant predictor of responsibility judgments [$b = 13.7$, $t(465) = 4.35$, $p < 0.001$]. The regression coefficients for the ring [$b = 1.2$, $t(465) = 0.82$, $p = 0.414$] and the score [$b = 0.5$, $t(465) = 1.41$, $p = 0.157$] variables were not significant (see **Supplementary A** for the model predictions). We obtained similar results for the causation judgments. The regression coefficient was only significant for the number of dice [$b = 13.8$, $t(447) = 4.287$, $p < 0.001$]. The coefficients of the ring [$b = 0.5$, $t(447) = 0.31$, $p = 0.754$] and the total score [$b = -0.2$, $t(447) = 0.60$, $p = 0.548$] variables were not significant.

Finally, the results for the cases of failure, reflect those for the wins (see **Supplementary A** and **Supplementary Figure S1**). We find that the only variable that affects either judgment is the number of dice rolled – the variable that controls the success rate. The more dice were rolled (as a result of a better dart-throw), the lower was the ascription of causal responsibility to the agent for an eventual failure, and the lower was the extent to which people thought the agent caused the failure (see **Supplementary A** for full report of statistical tests).

Discussion

As in previous studies (Gerstenberg et al., 2018), participants judged agents whose actions yielded higher success rates to be more responsible in case of success, but less responsible in cases of failure. The exact same pattern is shown for judgments of causal strength. Both the participants' responsibility and causation judgments support the robust-causation hypothesis rather than the skill-only hypothesis. We reasoned that an agent's variation in skill would correspond to more localized dart throw gradients (around the dart-board center, i.e., the 6-ring), and thus would result in more throws ending in even (2, 4, 6) compared with the odd (1, 3, 5) rings[8]. There was no such difference in either the causal strength nor the causal responsibility judgments. On the other hand, the robust causation hypothesis correctly predicts that people's judgments of causation and responsibility are a function of the number of dice rolled, and not simply the closeness of the dart to the center of the board. The total score (which involves the luck of the dice throw) also did not affect the judgments either (as long as it was at least 6), when holding the

[8] One may contend that a perfectly skilled and rational agent should be indifferent to differences between rings that are equivalent in the number of dice deployed. Such an agent, however, should always hit rings 5–6, and thus is inconsistent with most cases shown in the experiment. A more plausible interpretation of skill is to accept some unavoidable noise in the outcome and thus aim as close to the center as possible.

number of dice constant, indicating that judgments were affected by the agent's action, rather than by resultant luck (Nagel, 1979; Gerstenberg et al., 2010).

Nevertheless, one might argue that there is a simpler framework that can account for this pattern of results, which is independent of the notion of robustness. The idea is that as we increase the number of dice being rolled, we also increase the success probability, and an increased probability of success may lead to increased judgments of causal responsibility and causal strength (Suppes, 1970; Cheng and Novick, 1992; Spellman, 1997). Note, however, that probability-raising and robustness are not independent, and that one very typical way in which one increases the success probability of an action is by making it more robust. However, it is possible to tease apart robustness and probability-raising empirically. In the following experiments, we test the robustness hypothesis while keeping the probability of success fixed across conditions. To do so, we will contrast two types of actions whose success rate is the same, but which vary in their robustness. In Experiment 2, we use an animated soccer scenario, while in Experiments 3 and 4, we use a vignette based on an analog of Ellsberg's ambiguity paradox.

EXPERIMENT 2

In this experiment, we hold probability-raising constant and manipulate robustness in a more naturalistic setting – a soccer set-up based on video animations. Since we obtained in the previous experiment parallel results with the causal-strength and the responsibility measures, and since our main interest is the responsibility that agents have for the effects of their action, we explicitly probed here the responsibility and praiseworthiness of the agent. Participants view an animation with soccer players taking free-kicks, and with three defenders that form a (slightly moving) defensive wall in front of the goal (see *demo* link in the Materials section). The soccer players vary on two orthogonal factors – the probability of scoring a goal and the execution strategy (both manipulations were carried out within participants). In order to establish generality, we carried out two versions of this experiment, in which we manipulated the success probability by experience (Experiment 2a) and by description (Experiments 2b and 2c)[9]. There were two strategies: (i) The non-robust strategy of shooting directly into the wall (this may result in a goal, if the defenders happen to accidentally move out of the way). (ii) The robust strategy of shooting in a curved trajectory around the wall (if successful, this strategy is robust to the location of the defenders). Since a good execution based on the second strategy might be perceived as more difficult to achieve, we made it explicit that the probability of scoring a goal (for the robust and non-robust players) were identical. Participants evaluated the responsibility of four players (2 × 2 design) in scoring a specific goal in a free-kick for their team in a decisive match.

As predicted by both robustness and the probability-raising theory, we expected that participants would rate higher the responsibility of the players who have a higher scoring record, compared with players with a low scoring record whose goals may be perceived as "lucky" (cf. Johnson and Rips, 2015; Gerstenberg et al., 2018). As predicted by the robustness hypothesis, we also expected that responsibility judgments would be higher for players who bend their shots around the wall, even when the success rate is equated. Players who successfully bend their shot scored their goal in a way that is not dependent on the particular background circumstances (and thus are less "lucky").

Method

Participants

Twenty-two participants (4 female, 18 males; mean age = 28.67, *SD* = 8.45) were tested in Experiment 2a – the *experience*-condition. These participants were either Tel Aviv University students (16 participants) that received 15 min credits for participating in the lab, or volunteers (6 participants) who ran the experiment on remote computers via a link to the same Qualtrics site. Twenty participants (1 female, 19 males; mean age = 33.8, *SD* = 12.8) were tested in Experiment 2b (*description*-condition in which the success rate of each player was verbally stated). These participants were recruited via "Hamidgam project" (the Israeli equivalent of Amazon Mechanical Turk) and received a payment that was equivalent to 0.41$ for participating. Three additional participants were excluded because they reported internet connectivity problems (2), or because they didn't watch or play soccer at least once a year (1). In Experiment 2c (replication of Experiment 2b), we tested 46 participants (46 males, mean age = 29.09, *SD* = 4.47) via "i-Panel."[10] These participants received a payment that was equivalent to 0.27$. All participants across all three studies reported watching or playing football at least once a year.

Materials and Procedure

Participants took part in the lab or remotely, using their computer. After a short instruction, six sample free kicks for each player (experience condition) or a table with the player's success rates and kick-style (description condition; see **Supplementary B**) were shown. Video-clips were shown and ratings were made using the Qualtrics platform. The video clips may be accessed here: https://github.com/guygrinfeld/Responsibility-and-Robust-Causation-Experiments/tree/master/videos-sample.

After watching each player (or reading information about each player), a video-clip of a successful free-kick was shown (see **Figure 3**) and subjects were asked to evaluate: "How responsible and praiseworthy is the player for this goal?" on a scale from 0 (not responsible at all) to 100 (has full responsibility). Four players were presented and rated sequentially, according to the same procedure. Each player had a different color shirt to help

[9] Experience based manipulations are more similar to real life, but they could be subject to biases in priors. The description-based manipulation explicitly states the success probability rather than relying on participants' inferences from observation. Experiment 2c is a replication of Experiment 2b, which was carried out during the review of the manuscript.

[10] We replicated Experiment 2b with a sample-size of 46 which was based on a power analysis that resulted in 80% power for the robustness effect of Experiment 2b.

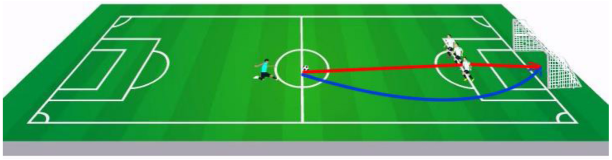


FIGURE 3 | Illustration of the robust (curved, blue line) and non-robust (straight to the wall, red line) free kicks. In the animation, the defenders are moving, such that their locations are difficult to predict (see demo: https://github.com/guygrinfeld/Responsibility-and-Robust-Causation-Experiments/tree/master/videos-sample).

differentiate the players. The four players had the following characteristics:

(1) Robust player (curved trajectory), low success rate: In Experiment 2a, this player shoots the ball in a curved trajectory and succeeds in two out of six free-kicks (i.e., 1/3 success rate). Successful free-kicks were second and fourth. In the description condition, the success rate was stated as 30% in Experiment 2b and as 1/3 in Experiment 2c.
(2) Robust player (curved trajectory), high success rate: This player scores in four out of six free-kicks (i.e., 2/3 success rate). Successful free-kicks were first, third, fifth and sixth. In Experiment 2b, the success rate was stated as 60% and in Experiment 2c as 2/3.
(3) Non-robust player (straight trajectory), low success rate: In Experiment 2a, this player shoots the ball straight at the wall and succeeds in two out of six free-kicks (1/3 success rate). Successful free-kicks were second and fourth. In Experiment 2b, the success rate was stated as 30% and in Experiment 2c as 1/3.
(4) Non-robust player (straight trajectory), high success rate: This player scores four out of six free-kicks (2/3 success rate). Successful free-kicks were first, third, fifth, and sixth. The success rate was stated as 60% in Experiment 2b and as 2/3 in Experiment 2c.

The order of four players was randomly assigned to each participant.

Results

The responsibility judgments are shown in **Figure 4**. Three 2 × 2 ANOVAs with robustness (robust vs. non-robust) and success-rate (1/3 vs. 2/3) as within-subject factors were carried out. In Experiment 2a, we find main effects of robustness and of success rate. The participants rated the players that took robust shots as more responsible for the goal than those that took non-robust shots, $F(1,21) = 13.01$, $p = 0.002$, $\eta_p^2 = 0.408$. Participants also rated players who had a high success rate more responsible for their goals, compared to players who had a low success rate, $F(1,21) = 14.10$, $p = 0.001$, $\eta_p^2 = 0.426$. There was no significant interaction between robustness and success rate, $F(1,21) = 0.12$, $p = 0.733$, $\eta_p^2 = 0.006$.

In Experiment 2b (description), participants also rated the robust players as more responsible for the goal than those that took non-robust shots, yet this effect was only marginally significant, $F(1,19) = 3.06$, $p = 0.095$, $\eta_p^2 = 0.127$. As in Experiment 2a there was a main effect for success rate, $F(1,19) = 6.38$, $p = 0.020$, $\eta_p^2 = 0.233$, and no interaction between robustness and success rate, $F(1,19) = 0.26$, $p = 0.617$, $\eta_p^2 = 0.012$. Because robustness was only marginally significant in the description condition, we performed a replication study with a larger number of participants (Experiment 2c; see Footnote 10). We replicated our findings. Participants rated the robust players as more responsible for the goal than those that took non-robust shots, $F(1,45) = 4.24$, $p = 0.045$, $\eta_p^2 = 0.086$. Participants also rated the more successful players as more responsible, $F(1,45) = 8.19$, $p = 0.006$, $\eta_p^2 = 0.154$, with no interaction between these two factors, $F(1,45) = 1.42$, $p = 0.240$, $\eta_p^2 = 0.030$.

Discussion

The results indicate that both success rate and robustness (as operationalized by the action-outcome contingency being dependent on background conditions), independently affect causal responsibility ratings. Note that while the effect of robustness appears larger in the experience condition[11], it is also present in the description condition. For the experience-based condition, one may argue that the participants make inferences on the success probabilities that are subject to bias in the priors, resulting in larger success rates for the curved (thus impressive) compared with the straight strategy kicks. Such biases could generate a robustness effect due to a success-rate artifact. Such a memory bias effect, however, is less plausible in the description condition, in which the success rates are explicitly stated (see Experiment 4 for an explicit test showing that the participants showed accurate memory of stated success rates).

The results indicate that the robustness of an action (the way the player takes the free-kick) affects the attribution of responsibility for its outcome (the goal), independent of the success rate. The results also consistent with the probability-raising principle, and indeed robust actions are typically more likely to succeed (but see Experiments 3–4 and Vasilyeva et al., 2018, for situations in which this is not the case). Both of these results could be understood to result from the negative influence of luck on the degree of responsibility, and from the idea that goals by players who shoot through the wall are more likely to have resulted from luck (cf. Gerstenberg et al., 2018).

There is, however, an alternative interpretation for the robustness effect. Accordingly, one may argue that the robust and successful kicks involve more skill, and therefore, it is a skill and not robustness *per se* that affected participants' judgments. While the skill hypothesis was not supported in Experiment 1, Experiment 3 will control both probability-raising and skill.

EXPERIMENT 3

In this experiment, we further test for robustness using a design that controls both probability-raising and skill. Our experimental

[11] It is possible that the larger effects in Experiment 2a are due to most participants being tested under more controlled lab conditions rather than online.

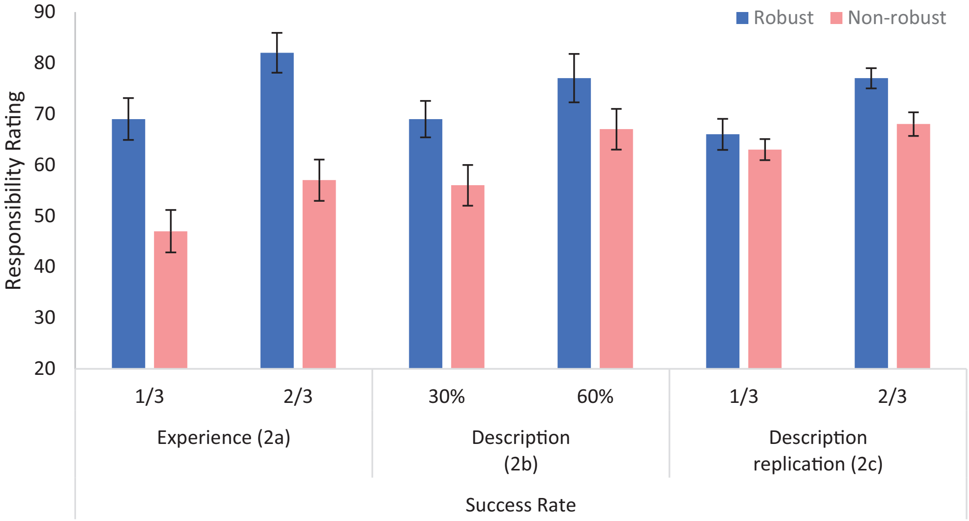


FIGURE 4 | Judgments of causal responsibility, as a function of success rate and robustness, in Experiment 2, for the Experience (left panel, *n* = 22) and for the Description (middle panel, *n* = 20; and right panel, *n* = 46) conditions. Error bars indicate within-subject ±1 standard error of the mean.

TABLE 2 | An adaptation of the Ellsberg paradox.

	Red balls	Black balls	Yellow balls
Urn 1	30	10	50
Urn 2	30	50	10

The values indicate the number of differently colored balls in two different urns.

set-up is inspired by the Ellsberg paradox (Ellsberg, 1961*;* see also Vasilyeva et al., 2018). Consider an agent that is faced with a lottery in which a ball is randomly selected from one of two urns (**Table 2**). In Urn 1 there are 30 red, 10 black, and 50 yellow balls while in Urn 2 there are 30 red, 50 black, and 10 yellow balls. The agent has no control over which urn is chosen; the urn is selected by another person. The agent also does not know the probability with which this person chooses Urn 1 versus Urn 2 and can bet either on red or on black. Which color should she bet on? The typical Ellsberg paradox result, is that when faced with such choices, agents prefer to bet on red (which has a definite 1/3 probability of a win) than on black (whose win probability is not determined and can be anything between 1/9 and 5/9). Here, we don't focus on preferences, but rather on how observers judge the causal responsibility of agents who achieve a goal by taking a robust or a non-robust action.

Note, that for this situation, the outcome of the action (bet-red or bet-black) depends probabilistically on both the ball selected and on the background condition (the other person selecting Urn 1 or Urn 2). The bet-red action is thus more robust than bet-black, because its win probability is stable across background conditions (Urn 1 vs. Urn 2). By betting on red, one's probability of success is rendered independent of the urn selection, whereas betting on black entails that one's probability of success depends on the urn chosen by the other person.

According to the robustness hypothesis, we predicted that participants would judge that an agent who wins as a result of a robust bet is more responsible than one who wins as a result of an unstable bet. Note that here there is no difference in the skill that the execution of the strategy requires. In addition, we also manipulated whether the background condition (Urn 1 vs. Urn 2) was decided by another person or by a computer. We hypothesized that if the background circumstance involves another agent who acts intentionally, the robustness effect will be enhanced, compared with a background circumstance that involves a non-intentional mechanism (Lombrozo, 2010). One possibility, for example, is that the presence of an agent (as a background condition) makes this background condition more salient.

To make the setting more realistic for our student participants, we framed the task in an exam setting. In particular, participants were asked to rate a candidate's responsibility for exam successes, with relation to his/her preparation for this exam. Participants read about six candidates, each of whom prepared differently in the way they allocated study time to the potential exam-topics (studied both topics or only one of them) and with regards to the total time they studied (duration of 3 or 5 days). The exam topic was picked randomly by the computer or by another agent. In this experiment, we were interested in responsibility for success in the exam. However, we added two additional filler candidates who failed the exam, in order to make the test more ecologically valid (it is unlikely that all candidates would be successful).

Method

Participants

Twenty Tel Aviv university students (17 females, 3 males; mean age = 22.9, *SD* = 2.0) took part in return for 15 min credit points needed in their BA requirements. Subjects spoke Hebrew as their mother tongue (17), or had at least advanced Hebrew reading capabilities (3).

Materials and Procedure

The experiment was run in lab, and presented via Qualtrics. First, participants read the following short text that introduced a hypothetical hiring procedure (exam topics and success rates were bolded and colored to ease the tracking of details; see **Supplementary C**):

"Assume that Google is recruiting new employees who need to pass a knowledge and ability exam. The exam questions this year may involve one of two possible topics: Algorithms or Cryptography (Google informs the candidates about the possible topics 1 week before the exam). Assume also that a candidate who is good at programming and did a BA in Computer Science has a 30% chance to pass the exam without any special preparation, no matter what topic is tested. However, if the candidate studies for the exam, her/his chance to pass it will increase as follows:

- A candidate who studies for 3 *days* on both topics (sharing time between them), will pass the exam with 50% chance if s/he gets a question about Algorithms, and also a 50% chance to pass if asked about Cryptography.
- A candidate who studies for *3 days* but chooses to learn only one topic, will pass with 70% chance if tested on this topic, but remains at 30% if asked about the other topic.
- A candidate who studies for *5 days* on both topics, will pass the exam with 60% chance if s/he gets a question about Algorithms, and also a 60% chance to pass if asked about Cryptography.
- A candidate who studies for *5 days* but chooses to learn only for one topic, will pass with 90% chance if tested on this topic, but remain at a 30% chance if asked about the other topic.

Next, participants were presented with information about a number of candidates for the latest Google-exam (see **Table 3**), all having "very similar intellectual abilities and programming skills, as reflected by their BA record, but differing in the way they prepared for the exam, and on the circumstances that determined the exam topic." The participants were asked to evaluate the responsibility of each candidate for passing or failing the exam on a 1–100 slider scale bar (see **Supplementary C**). The candidates (and their evaluations) were presented sequentially (not in table format), and participants were told that they can take as long as they need and that they are allowed to look back and compare previous judgments.

Participants were told that "in the morning of the exam, the exam-topic of *Cryptography* was randomly chosen by the computer software." After answering all of the eight evaluations, the participants were asked to make the same (A–H) evaluations again with one difference, which involved the manner in which the exam-topic was selected. Instead of having the topic randomly selected by the computer-software, participants were told that "in the morning of the exam, the topic of *Algorithms* was randomly chosen by computer software, but the head of recruitment decided not to let the computer determine the exam-topic and switched it to *Cryptography*."

The reason for the second set of evaluations was twofold. First, we wanted to test if the effects (of success rate and of robustness) are stable. Second, we wanted to test if the robustness effect is modulated by the presence of another agent, who is involved in the setting of the background conditions that, together with the candidate's action, determine the action's success (a type of responsibility dilution). For each exam candidate, the participants were asked to rate "to what extent is the candidate responsible for his success/failure in the exam?" (A screenshot of the materials is presented in the **Supplementary C**).

TABLE 3 | The candidates presented for judgment in Experiment 3.

Candidate	Robust
A studied for 3 days on both *Algorithms* and *Cryptography* and passes the exam.	+
B studied for 5 days on both *Algorithms* and *Cryptography* and passes the exam.	+
C studied for 3 days only on *Cryptography* and passes the exam.	–
D studied for 5 days only on *Cryptography* and passes the exam	–
E studied for 3 days only on *Algorithms* and passes the exam.	–
F studied for 5 days only on *Algorithms* and passes the exam.	–
G studied for 3 days on both *Algorithms* and *Cryptography* and fails the exam.	+
H studied for 5 days only on *Cryptography* and fails the exam.	–

Analysis

We focus on a number of contrasts, based on the candidates who passed the exam. We focus on the candidates that succeed in their exam, because our theory of robust causation depends on the agent taking an action that is intended to bring about an event (Woodward, 2006; Usher, 2018). Thus, robustness manipulations should be tested on events that match the agent's intention (success cases) and not on events that do not match (failures); but see further discussion for the case of failure in Experiment 4.

First contrasting robust candidates (A and B) with non-robust candidates (C and D; see **Table 3**) provides an estimate of the robustness effect. Second, contrasting candidates A and C, who studied for 3 days, with candidates, B and D, who studied for 5 days, provides an estimate for the effect study-duration (3 vs. 5 days). Third, comparing candidates C and D, who passed by studying only the selected topic, with candidates E and F, who passed by studying the topic that was not selected, provides an estimate of the effect that the match between the topic studied and the one selected makes for non-robust type actions (this match affects the success rate, conditioned on the background condition that was active). For example, if one studies the topic that was probed, the success rate should be inferred to be higher, and we predict that this will affect the responsibility ratings for success in the two cases. Fourth, the difference between the robustness effect in the computer-condition and in the "head of recruitment" condition reveals whether the robustness effect is modulated by the presence of an agent.

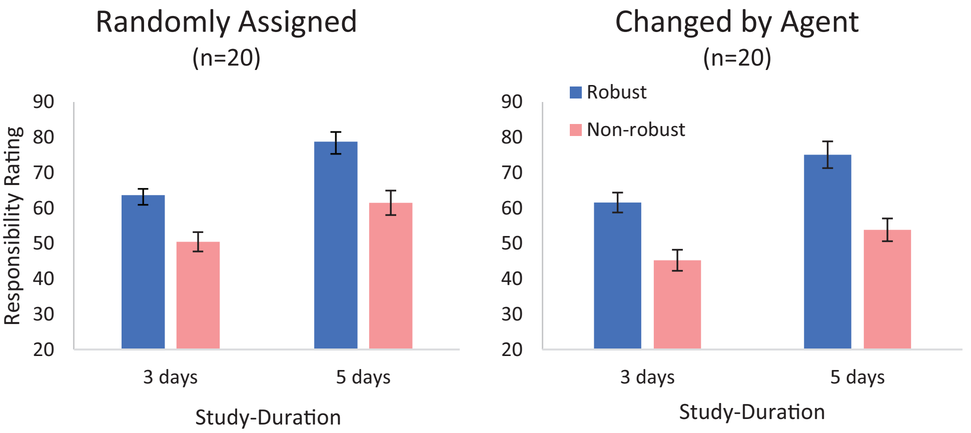


FIGURE 5 | Ratings of responsibility (cases A–D, in **Table 3**) in Experiment 3. Both the study-duration and the robust/stable action received higher ratings. Error bars indicate within-subject ±1 standard error of the mean.

Results and Discussion

Planned comparisons provided significant differences for all the variables above. Specifically, there was a significant robustness effect, where responsibility ratings of robust candidates A and B (M = 69.82, SD = 17.21) were higher than of non-robust candidates C and D [M = 52.80, SD = 19.00; $F(1,19) = 12.69$, $p = 0.002$, $\eta_p^2 = 0.400$], and a significant study-duration effect, where responsibility ratings of the three-learning-days candidates A and C (M = 55.28, SD = 15.70) were lower than of five-learning-days candidates B and D [M = 67.34, SD = 14.27; $F(1,19) = 69.29$, $p < 0.001$, $\eta_p^2 = 0.785$; see **Figure 5**]. This is consistent with the 2 × 2 × 2 within subjects ANOVA (on the A–D items), which resulted in three main effects (robustness, success-rate, and agent-framing) but no significant interactions. The "head of recruitment" framing reduced responsibility judgments, $F(1,19) = 5.61$, $p = 0.027$, $\eta_p^2 = 0.228$. However, while this effect was numerically larger for the non-robust (C and D) than for the robust (A and B) candidates, this difference did not reach statistical significance, $F(1,19) = 0.79$, $p = 0.384$, $\eta_p^2 = 0.040$.

Finally, we examine the effect that the difference in match (between the topic studied and the one selected) makes for non-robust type actions (C and D vs. E and F). Planned comparisons revealed that participants rated the match cases higher (studied *Cryptography* and exam-topic was *Cryptography*) than non-match cases [studied *Algorithms* and exam-topic was *Cryptography* $F(1,19) = 11.37$, $p = 0.003$, $\eta_p^2 = 0.374$].

The results of this experiment confirmed most of our predictions. First, as predicted by both robustness and probability raising, participants gave higher responsibility ratings for exam success to candidates who had a higher chance of success as a result of studying more. However, as predicted by robustness alone, participants rated robust candidates higher, who studied both topics, and thus their success was less dependent on background conditions.

Nevertheless, it is possible to query the type of judgments that the participants made. While we formulated this in terms of the "to what extent is the agent responsible for success/failure in the exam," one may also wonder whether participants distinguish between this and mere causal strength evaluation. Indeed, many experiments (including our Experiment 1) obtain similar results with causal strength and causal responsibility judgments. Our final experiment aims to contrast between these two measures and also to extend the judgments from cases of success to both success and failure.

EXPERIMENT 4

In Experiment 4, we aimed to replicate the results of Experiment 3 (validating memory of the success rates) and to contrast judgments of responsibility and judgments of causal strength. Although often, these two types of judgments have parallel effects, we expect these judgments to come apart in this specific setup. Compare, for example, candidates A and C (see **Table 3**), both of whom succeeded in the exam after having studied the same amount, but with A having divided the study among the two topics, while C having studied, only the topic that was tested. Following Chockler and Halpern (2004), we proposed that when judging the extent to which each candidate is responsible for the exam's success/failure, participants will take the *epistemic perspective* of the candidates at the time they made the action. On the other hand, when asked to evaluate the causal strength by which the action caused the effect, we expect participants to take an objective perspective, which includes the actual background circumstances. Indeed, in the actual situation in which the exam topic was chosen for which the non-robust candidate studied, the non-robust candidate has a greater success contingency than the robust-candidate.

In addition, we wanted to extend the range of cases to include cases of failure. For the case of responsibility, we do not make a specific prediction on how robustness (as expressed by studying a single or two topics) will affect the responsibility of failure (This is because failures do not satisfy the intentional-match requirement in robust action, and the robust action is more stable in its prediction of both success/failure). However, we expect a dissociation between the effects of study-duration on causal strength and responsibility judgments. Consider the case of an agent studying for only one topic, which does not come up in the exam, resulting in exam failure. While judgments of causal strength should be invariant to how long the agent studied (as the amount of studying the wrong topic should not affect the contingency with exam success/failure), judgments of causal responsibility are expected to decrease with study duration. Indeed, if judging responsibility depends on adopting the epistemic perspective of the agent, the actual background circumstances (topic mismatch) should not be assumed, and therefore, the more an agent studies for the exam, the more responsible she is for success (and less for failures), independent on whether the topic matches or not.

The experiment was identical to Experiment 3, except for a few modifications. First, we removed the head of recruitment vs. computer condition (we kept the computer framing only), and we included four candidates (A–D) that succeeded in the exam, and four candidates who failed (two who studied both topics, and two who studied the wrong topic). Second, we manipulated the type of rating (causal responsibility vs. causal strength) between participants in order to prevent a carryover between the two types of judgments. This allowed us to test the predicted dissociation between causal strength and responsibility judgments in cases of success, based on adopting the agent's epistemic perspective in the latter. Finally, we also included a post-test memory check, in which we asked participants about the success rates of the various candidates, in order to ensure that participants based their judgments on the data we provided.

Method

Participants

Sixty students at Tel Aviv University (30 in each condition[12]; 23 females, 37 males; mean age = 22.5, *SD* = 1.6) participated in this study in return for 15 min credit points.

Materials

The framing of the story was identical to Experiment 3. The eight candidates presented for evaluation are shown in **Table 4**.

Procedure

Responsibility judgments were assessed in the same way as in Experiment 3. In the causal strength condition, participants were asked: "To what extent did the study of the candidate cause the outcome in the exam?" As a memory check, after judging the candidates, participants were asked to fill in a table with success rates of the various candidates.

Analysis

Based on the predictions we outlined, we carried out 2 × 2 ANOVAs for passing candidates with factors of robustness and study-duration, separately for each judgment type (responsibility vs. causation). While cases of failure do not satisfy the intentional match criterion above, we also report the responsibility for these judgments, and we carry out a similar 2 × 2 × 2 ANOVA for the cases of failure.

TABLE 4 | The job-candidates presented for judgment in Experiment 4.

Candidate	Robust
A studied for 3 days on both *Algorithms* and *Cryptography* and passes the exam.	+
B studied for 5 days on both *Algorithms* and *Cryptography* and passes the exam.	+
C studied for 3 days only on *Cryptography* and passes the exam.	–
D studied for 5 days only on *Cryptography* and passes the exam	–
E studied for 3 days on both *Algorithms* and *Cryptography* and fails the exam.	+
F studied for 5 days on both *Algorithms* and *Cryptography* and fails the exam.	+
G studied for 3 days only on *Algorithms* and fails the exam.	–
H studied for 5 days only on *Algorithms* and fails the exam.	–

Results and Discussion

The post-experimental memory test showed that the participants remembered well the success rates of the eight candidates that they were required to rate, as indicated by the post-experimental memory test (see **Supplementary Figure S2** and **Supplementary C**). We now turn to the ratings of causal responsibility and of causal strength.

Causal Responsibility

For successful candidates (A–D in **Table 4**), we replicated the results of Experiment 3. There were two main effects, for robustness [$F(1,29) = 8.50$, $p = 0.007$, $\eta_p^2 = 0.227$] and for study-duration [$F(1,29) = 34.62$, $p < 0.001$, $\eta_p^2 = 0.544$], respectively. As shown in the upper-left panel of **Figure 6**, participants gave higher ratings to the robust study candidates and also to candidates who studied longer. There was also an interaction between these two factors [$F(1,29) = 5.46$, $p = 0.027$, $\eta_p^2 = 0.158$]. However, the simple effects of robustness were significant at both study-duration conditions [for 3 days, $F(1,29) = 7.09$, $p = 0.012$; for 5 days $F(1,29) = 8.94$, $p = 0.006$].

For the failed candidates (E–H in **Table 4**; upper-right panel in **Figure 6**), we obtained a main effect of robustness. Participants rated the candidates who studied in a robust way (both topics) as less responsible for their failure ($M = 46.89$, $SD = 25.97$) than the ones who studied in a non-robust way [on the topic that was not chosen, $M = 67.23$, $SD = 23.52$; $F(1,29) = 18.70$, $p < 0.001$, $\eta_p^2 = 0.392$]. We also found a main effect of study-duration [$F(1,29) = 14.36$, $p = 0.001$, $\eta_p^2 = 0.331$] and an interaction between robustness and study-duration [$F(1,29) = 17.00$, $p < 0.001$, $\eta_p^2 = 0.370$]. The duration of study reduced the responsibility for failing in the exam *only* for the candidates that studied both topics [$F(1,29) = 22.13$, $p < 0.001$, $\eta_p^2 = 0.433$].

[12] Based on the effect size in Experiment 3, this sample should allow a 95% power for replication of the robustness effect.

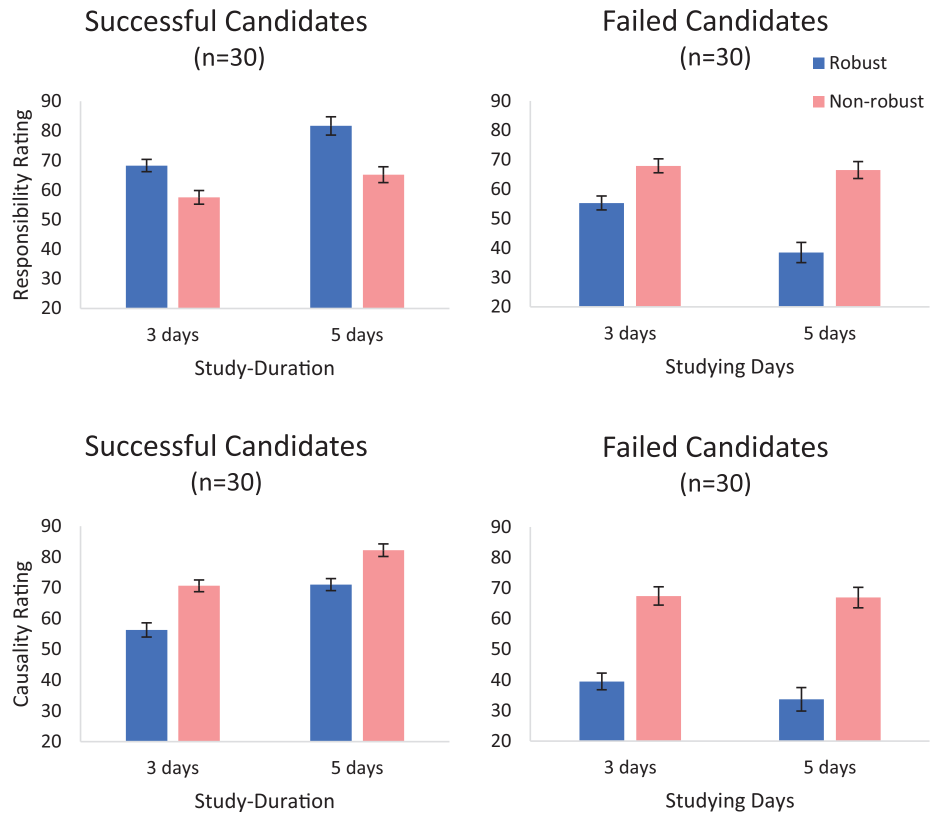


FIGURE 6 | Responsibility ratings (upper panels) and causal strength ratings (lower panels) for passed (left) and failed (right) candidates in Experiment 4. Error bars indicate within-subject ±1 standard error of the mean.

Causal Strength

For successful candidates, judgments of causal strength showed main effects of robustness and study-duration. As in the responsibility condition, the duration of study increased the ratings [$F(1,29) = 153.72$, $p < 0.001$, $\eta_p^2 = 0.841$; lower-right panel]. However, oppositely to the responsibility condition, here robustness strongly *decreased* causal strength ratings [$F(1,29) = 11.55$, $p = 0.002$, $\eta_p^2 = 0.285$]. Indeed, participants judged that a candidate who spent her time studying only the topic that was selected, caused the success in the exam to a higher degree ($M = 76.36$, $SD = 21.05$) than one who spent the same amount of time studying both topics ($M = 63.61$, $SD = 19.94$). Thus, judgments of causal-strength, but not of causal responsibility, appear to track the extent to which the action increased the probability of the outcome (conditional on the actual background conditions).[13] There was no interaction between study-duration and robustness [$F(1,29) = 1.62$, $p = 0.213$, $\eta_p^2 = 0.053$].

Finally, for failed candidates there was a main effect of robustness. Participants saw non-robust candidates who studied the wrong topic to have caused their failure to a higher degree ($M = 67.20$, $SD = 29.44$) than robust candidates who studied both topics [$M = 36.58$, $SD = 19.39$; $F(1,29) = 24.74$, $p < 0.001$, $\eta_p^2 = 0.460$]. Also, like for casual responsibility, the amount of study reduced casual strength [$F(1,29) = 4.75$, $p = 0.037$, $\eta_p^2 = 0.141$]. This reduction seems to be stronger in candidates who studied both topics, but this interaction (between robustness and study-duration) was not significant [$F(1,29) = 2.50$, $p = 0.125$, $\eta_p^2 = 0.079$].

GENERAL DISCUSSION

In four experiments, we tested if the extent to which agents are held causally responsible for the outcomes of their actions is affected by the robustness with which the action brought about the outcome (Woodward, 2006; Icard et al., 2017; Vasilyeva et al., 2018). In the first experiment, we manipulated robustness by explicitly increasing the number of background circumstances (possible dice outcomes) in which the action (dart throw) results in a win. We did this by contrasting actions (dart throws) that result in one, two or three dice rolls, depending on the ring number of the dart on the board (where success requires a sum of 6 on the dice). We found that agents whose dart throws result in more dice are seen as more responsible for success (and less responsible for failures). We observed the same

[13] See Endnote 4 in Vasilyeva et al. (2018), for results showing that judgments of causal strength track the extent to which the action increased the probability of the outcome (conditional on the actual background condition).

pattern of results for judgments of causal strength. The parallel effects of robustness on judgments of causal strength and causal responsibility are consistent with the responsibility-view, which sees causal attributions as mediated by normative attributions about what an agent should have done in a given situation (Sytsma et al., 2012). To achieve more dice rolls, one needs more skill. However, by comparing even vs. odd ring-numbers, we found that participants' judgments of causal responsibility and causal strength tracked robustness and were not merely affected by the skill of the player. These results are also consistent with *probability-raising* accounts according to which responsibility and causal strength judgments track the extent to which an action increased the probability of the observed outcome (Suppes, 1970; Cheng and Novick, 1992; Spellman, 1997; see also Kominsky et al., 2015; Icard et al., 2017, for recent studies showing that judgments of causal strength vary with the typicality of the cause and the background conditions).

Experiments 2–4 tested for effects of robustness while controlling for probability raising. Experiment 2 examined a soccer scenario in which strikers had two different ways of taking free-kicks. The non-robust action is to shoot the ball directly through the defensive wall. Such an action may result in a goal, depending on background circumstances that the agent does not control (the position and movement of the defending players). The robust action is to bend the ball around the wall; if well-executed, this action results in a goal in a way that depends less on background conditions (the exact location of the defenders do not matter). Because taking a well-executed curved kick is difficult, players may have similar success rates when employing the two strategies. We thus presented participants with animations of such hypothetical players, and independently manipulated success probability and robustness. To ensure that participants are not biased in their success probability assessment by the type of kick (due to prior expectations), we included a condition that stated the success probability, rather than leaving it for participants to estimate. For both description and experience conditions, we found that ratings of responsibility increased with robustness, even when the probability of success was kept constant. While these results support the robustness hypothesis, they are subject to an alternative explanation. In particular, it is possible that the ratings don't reflect considerations of robustness, but rather the inferred skill of the agent (cf. Gerstenberg et al., 2018). Experiments 3–4 addressed this issue using a scenario based on an Ellsberg-type design and using an exam-success setting (see also Vasilyeva et al., 2018, for a similar design).

In Experiment 3, the action (the way to prepare for the exam) did not vary in skill between the robust (split study time between both topics) and non-robust (study only one topic) action. Also, the overall success rate of the robust action was equated with that of the non-robust one, but the outcome of the non-robust action was more variable, depending on an external factor (selected exam topic). Hence, an account based on probability raising would not predict any differences in judgments. Note, moreover, that after a particular exam topic was selected the probability of success is now in favor of the non-robust case if the topic selected matches the one that the candidate prepared for. Like in Experiment 2, the results showed effects of both success rate and robustness. In particular, participants judged candidates who prepared for both topics more responsible for their exam success than those who only prepared for a single one, and were lucky in that this topic was chosen. Note also that judgments of responsibility tracked robustness even though, as stated above, the non-robust candidates actually had a higher probability of success given the lucky background.

Finally, in Experiment 4, we replicated the results of Experiment 3 under two important modifications. First, in addition to assessing causal responsibility, we also assessed judgments of causal strength. We predicted that the role of robustness would be different in judgments of causal responsibility versus causal strength, because the agent's epistemic perspective is more important for judgments of causal responsibility (in this case, we predicted that participants would adopt the agent's epistemic perspective). The results fully replicated the results of Experiment 3 in the responsibility/success case. Second, we examined how study-duration and study-type affect the responsibility and the causal strength judgments in cases of failures. Consistent with previous findings (Gerstenberg et al., 2018), participants judged agents who took robust actions that yielded more stable success rate, to be more responsible in case of success, but less responsible in cases of failure. Note that while this result is easy to motivate for study-duration (because it is positively correlated with success-rate and negatively correlated with failure-rate), it is less straightforward for study-type. Here the robust action, of studying both topics, is more stable with regards to both the success/failure events. In other words, while a robust action that resulted in success is more stable to changes in background conditions, so is a robust action that resulted in failure. These results are consistent with the idea that the effect of robustness (studying one vs. both topics in Experiments 3–4) is evaluated based on the stability of the contingency between the action and the successful outcome (the action's goal). We propose that for cases of failure, the robustness is derived from a negation of the goal-achievement: because the agent is more responsible (when doing A compared with B) if she succeeded in achieving her goal, she is less responsible (when doing A compared with B), in case the goal was not achieved. Finally, we find that for cases of failure, neither causal responsibility nor causal strength are affected by the study duration, for candidates who study only the wrong topic. This shows that study duration alone affects causal judgments only when it plays a role in the causal chain of events from action to outcome.

While we attribute the robustness boost of the responsibility judgments in Experiments 2–4 to the lack of causal dependency on background conditions (external to the agent), such as other players (Experiment 2) or the exam selection (Experiments 3–4), it is still possible to suggest that some of the participants are nevertheless responsive to some inferred agential trait, such as skill (Experiment 2), or study efficiency (Experiments 3–4). We ruled out skill as a mediator, in our analysis of Experiment 1, and our instructions attempted to eliminate it as a factor in Experiments 2–4. For example, we explained that the success rate of robust/non-robust players is the same (Experiment 2), and that all candidates are equal in their computer science knowledge and abilities (Experiment 3–4). Nevertheless, it is still possible that some participants may have inferred that a robust candidate (who studied both topics and succeeded) is more effective than a non-robust one (who studied the one topic that happened to be assessed in the exam). Accordingly, as suggested by Gerstenberg et al. (2018) an inferred agent trait (skill) could potentially mediate the effect that robustness has on judgments of causal responsibility. While future studies will be needed to test the possibility of further dissociating robustness from skill, we believe that this should not be viewed as a confound. Rather skill is probably a necessary feature of agents who exercise robust control (Usher, 2018).

The most novel result of this experiment, however, was the difference between the patterns observed in the causal responsibility and the causal strength condition (see **Figure 6**, left panels). While the causal responsibility of the agent increased with robustness (as the participants took the agent epistemic perspective), the causal-strength decreased with robustness, as the participants took a more objective perspective, assuming knowledge of the actual background circumstance. We believe that this dissociation (and deviation from the responsibility view; Sytsma et al., 2012), which is a rare one in the literature, was made possible by the specific Ellsberg-type design, which allowed us to dissociate between the objective contingency and the agent-based epistemic one. While future studies will be needed to test the possibility of further dissociating robustness from inferred skill in this type of design, we believe that in ecological conditions (e.g., Experiment 2) skill and robustness are associated, as skill is typically a feature agents need to deploy in order to exercise robust control (Usher, 2018). In the following, we discuss the implications of our results for normativity and their relation with other related studies.

Normativity

The results of Experiment 1 are readily understood from a normative perspective: rational agents should aspire to increase the likelihood of their desired outcomes. As discussed by Woodward (2006) robust actions (shooting a person in the heart) are more likely to achieve a goal (like the death of the victim), compared to non-robust actions (shooting the victim in the leg, which may or may not result in death), as the outcomes of such non-robust actions are likely to depend on background circumstances. Similarly, Woodward (2006) has argued that robustness is a critical difference that distinguishes between cases of causation by action and causation by omission or by double prevention (see Lombrozo, 2010 and Cushman and Young, 2011, for experimental studies showing that participants are sensitive to these differences in their causal judgments). Furthermore, Lombrozo (2010) has argued that moral judgments are affected by the stability of the causal relation to variations in background circumstances. More recently, Usher (2018) has argued that in order to achieve robust causation of actions over desired outcomes, agents deploy a teleological guidance control that is based on a means-ends strategies (cf. Heider, 1958). As agents do not have access to all information on background circumstances, they should attempt to act so as to make the outcome less dependent on such circumstances. In our Experiment 1, achieving a 3 dice roll, grants the agent with more opportunities to succeed, making her less dependent on chance. A similar situation obtains in Experiment 2, by attempting a curved-style free-kick, the agent takes an action whose outcome is less dependent on circumstances beyond her control.

In Experiments 2–4, we clarified to our participants that the probability of success of the robust and non-robust action is the same. In Experiment 2, robustness to background circumstances was balanced by the difficulty of executing such an action. In Experiments 3–4, were inspired by Ellsberg scenario (Ellsberg, 1961) that let us keep the probability of success fixed but to vary the robustness. Still, we find that people evaluate the agent as more responsible for the outcome of her action, in the case of robust action. The normativity of this judgment, thus, requires a special discussion.

Our conceptualization of robustness, via a count of background circumstances that enable an intended event was proposed by Woodward based on a number of conceptual considerations, such as invariance (Woodward, 2003, 2006). This conceptualization also has the advantage that it does not require access to probabilities of the background circumstances (which are often difficult to access). Usually, robustness as measured by this count definition, correlates with the success probability as in our Experiment 1, however, robustness and probability raising can also stand in opposition. For example, one may contrast an action that produces an effect in 10 background conditions with a small probability (say, 5% in each) with another that produces the effect in a single background condition (but with a higher probability, of say, 90%). It is beyond our aim to make either normative or empirical claims about what is expected in such special situations. Our Experiments 2–4, kept the total success probability fixed while varying the count-type robustness. Thus, we believe they support the more modest conclusion, that once the success probability is fixed, the count-measure of robustness affects the judgments of causal responsibility. Future investigations will be required to examine tradeoffs between success probability and count-measures of robustness.

What our Experiments 3–4 show is that the preference for the stable alternatives (those whose success rate does not depend on factors that are not known, also labeled as *ambiguity-aversion*; Ellsberg, 1961), is also reflected when we judge agents who take robust actions (in the sense above) as higher in causal responsibility. We believe that the reason for this is the fact that the success of the non-robust action appears lucky (see also Gerstenberg et al., 2018), as it depends more on other agents or circumstances. In Gerstenberg et al. (2018), the contrast between agent-bearing responsibility actions (for which the agent gets high credit) and lucky ones (for which she gets less credit) was made via the contingency between the action and the outcome. Here, we kept this contingency constant (or even reduced it in the case of robust actions compared to non-robust matched actions, Experiments 3–4), but we manipulated the presence of non-agent background conditions. Thus, consistent with Woodward's theory of robust causation (2006), lucky actions are those in which the background conditions contributed significantly to the outcome, and thus, non-robust actions receive lower responsibility, reflecting a type of diffusion of responsibility among multiple causes (Lagnado et al., 2013). Indeed, in previous studies, Lagnado et al. (2013) showed that when the number of agents that disjunctively contributed to an event increases, the judged responsibility of each agent is reduced. Since our non-robust actions allow other agents (or factors) to contribute to the production of the intended event (note that this may involve a contribution by non-acting, as for the defenders who miss blocking the ball shot through the wall), we can think of their effect on judgments of responsibility as a type of responsibility diffusion. The robust actions, on the other hand, are such that they screen-off the intended event from the impact of other agents or background circumstances and thus, they satisfy a robust sufficiency criterion.

Finally, another advantage of robustness is that robust causal setups have the advantage that the causal Markov condition[14] is preserved on the level of overall categories. In contrast, in setups with subcategories with the same causal structure but different causal strength, the Markov condition does not hold globally, resulting in distorted judgments of correlations and causal relationships (von Sydow et al., 2016; Hebbelmann and von Sydow, 2017; cf. Hagmayer et al., 2011).

Relation to Other Work and Alternative Theories

In a recent paper, Vasilyeva et al. (2018) reported that people's judgments about causal generalizations and causal explanations are sensitive to the stability of these relations, even when probability-raising is controlled. In their studies, participants were presented with descriptions or contingency tables for a potential causal relation, and were then asked to indicate the degree to which they endorse a causal explanation (or causal generalization) for the situation described. For example, in studies 1 and 2, participants were presented with contingency tables for fictional lizard-like species (Zelmos), which either did or did not eat yona-plants (the action) and either did or did not get sore antennas (the effect). These tables included a moderating variable (drinking salty/fresh water) that varied or did not vary the relationship between the action and the effect. The presence of the moderating variable that affected the action-effect relationship reduced the degree of causal endorsement, even though the average causal strength across the moderating variable was the same. In their study 3, a similar result was obtained for the endorsement of a causal relation between people taking a vitamin and the effects on bone density, with gene-type as a moderating variable.

While these findings parallel our results from Experiment 3, there are a number of important differences, and thus we believe that the two approaches complement each other. The central difference is that while our experiment was designed to assess people's judgments of the extent to which an agent's action was causally responsible for bringing about an outcome they intended, Vasilyeva et al. (2018) assessed people's endorsement of causal relations and of causal explanations between (type or token) events.

For example, in one of their experimental conditions, which is most similar to ours, after being presented with the background and the contingency tables, participants were told: "Your assistants select one of the zelmos with sore antennas from your second experiment. They call him Timmy. During the experiment, Timmy has eaten yonas. *You do not know whether Timmy drank fresh water or salty water during the experiment*. How much do you agree with the following statement about what caused Timmy's sore antennas? *Eating yonas caused Timmy's antennas to become sore*" (Vasilyeva et al., 2018, **Table 2**, p. 8).

Compare this with our scenario in Experiments 3 and 4, where participants evaluated the causal responsibility that the agent's action (type of study) has for the outcome (success/failure in exam), under conditions that differ in sensitivity to an external circumstance (question chosen by the computer or by another agent). There are two important differences. First, as we formulate this at the level of agents taking an action toward a goal, we can probe the causal responsibility of the agent for the outcome of the action and contrast it with the causal-strength (it would make little sense to ask "how responsible is the Zelmo for getting sore antennas" in this context); agent-responsibility requires a set of minimum epistemic conditions, such as the agent foreseeing the potential consequences of her actions (or being in a state where she is expected to do so), which are in place in our case. While one may ask instead about the causal responsibility between the events ('eating yonas' and 'having sore antennas'), we point below to an important difference.

Second, there is an important epistemic difference. In the 'zelmos sore-antenna' case the reduced causal endorsement of the causal relation in the non-stable condition is conditioned on lack of knowledge: participants did not know about the state of the moderating variable ("*You do not know whether Timmy drank fresh water or salty water during the experiment*"). In our Experiment 3–4, on the other hand, participants knew the state of the background variable (the exam topic selected). In contrast to Vasilyeva et al. (2018), we obtained an increased degree of causal strength (in the non-robust condition), showing that when such information becomes available, the participants rely on it in their causal strength judgments, and they do not adopt the agent's epistemic perspective [see Endnote 4, in Vasilyeva et al. (2018), for a similar result]. Both of these results are normatively reasonable, as it makes sense to attribute increased causal strength to a causal relation that has a stronger contingency (our Experiment 4, and results reported in Vasilyeva et al., 2018, Endnote 4), and also to feel uncertain of the causal relation (Vasilyeva et al., 2018) given lack of knowledge on whether the sample belongs to a case that does or does not involve causal relation.

More importantly, we find that, even in a condition in which the causal contingency favors the non-robust action, causal responsibility judgments show a robustness effect: higher ratings for the robust actions. This provides a strong demonstration that robust actions confer more responsibility on an agent, even

[14] The causal Markov condition is a critical assumption of the Causal Bayesian framework (Pearl, 2000; Woodward, 2003)– whereby any variable in a causal model is conditionally independent of its non-descendants given its direct causes. While there is controversy about whether people uphold this condition in their intuitive inferences (see Sloman and Lagnado, 2015, for summary), it is a desirable property for normative inference.

if the actual state of the environment happens to be such that the actual probability of success is lower. We have argued that the difference between opposed effects of robustness on causal responsibility and causal strength (**Figure 6**), stem from the fact than in responsibility attributions, participants consider the epistemic state of the agent. We believe that, taken together, our studies and those of Vasilyeva et al. (2018), provide compelling and complementary evidence for the importance of robustness in the endorsement of causal responsibility relations between events and of causal explanation, and in judging the causal responsibility an agent has for the outcome of her action.

Further Implications and Future Research

We have focused here on judgments of *causal* responsibility. Future research is needed to clarify the normative aspect of stability in responsibility judgments, as well as its derivation from theoretical principles (e.g., Halpern and Hitchcock, 2015). Moreover, it has been argued that causal responsibility is a central component of legal and moral responsibility (Tadros, 2005; Moore, 2009; Lagnado and Gerstenberg, 2017; Usher, 2018). For example, it has been proposed that the degree of responsibility an agent has toward outcomes of her action depends on the teleological control that she deploys to achieve that effect (Lombrozo, 2010; Usher, 2018) and that differences in robust causation are the source of our feelings of a reduced responsibility toward manipulated agents (Deery and Nahmias, 2017; Murray and Lombrozo, 2017; Usher, 2018). Future research needs to test potential dissociations between robustness and skill and also examine how the attributions of responsibility change for teleological continual actions, in which the agent acts so as to carry out compensatory corrections needed to preserve a goal in the face of perturbations or interventions (Heider, 1958; Usher, 2018). Future research is also needed to examine potential distinctions between the causal responsibility of agents for the outcome of their intended actions (of the types we have examined here) and judgments of praise or blame.

In sum, robustness is an important but relatively under-explored causal concept (Woodward, 2006). Convergent evidence from our current studies, and also from Vasilyeva et al. (2018) using different experimental paradigms, show that robustness is itself a robust phenomenon in shaping people's causal judgments.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Experiment 1: Ethics Committee, UCL Experiments 2–4: Ethics Committee Tel Aviv University (1321253). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MU and DL conceived the idea. DL, MU, and TG designed Experiment 1. GG and MU designed Experiments 2–4. GG ran Experiments 2–4. GG ran all analyses. MU, DL, and GG wrote the manuscript. All authors read the manuscript and contributed to improving it.

FUNDING

This work was supported by ESRC grant ES/H041796/1 Causality and responsibility judgments to DL.

ACKNOWLEDGMENTS

We want to thank Jonathan Livengood for critical reading of the manuscript and to David Mason for help with designing materials for Experiment 1.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01069/full#supplementary-material

REFERENCES

Baron, J., and Ritov, I. (2009). Protected values and omission bias as deontological judgments. *Psychol. Learn. Motivat.* 50, 133–167. doi: 10.1016/s0079-7421(08)00404-0

Brickman, P., Ryan, K., and Wortman, C. B. (1975). Causal chains: attribution of responsibility as a function of immediate and prior causes. *J. Pers. Soc. Psychol.* 32, 1060–1067. doi: 10.1037/0022-3514.32.6.1060

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychol. Rev.* 104:367. doi: 10.1037/0033-295x.104.2.367

Cheng, P. W., and Novick, L. R. (1992). Covariation in natural causal induction. *Psychol. Rev.* 99, 365–382. doi: 10.1037/0033-295x.99.2.365

Cheng, P. W., and Novick, L. R. (2005). Constraints and nonconstraints in causal learning: reply to White (2005) and to Luhmann and Ahn (2005). *Psychol. Rev.* 112:694706. doi: 10.1037/0033-295X.112.3.694

Chockler, H., and Halpern, J. Y. (2004). Responsibility and blame: a structural-model approach. *J. Artif. Intellig. Res.* 22, 93–115. doi: 10.1613/jair.1391

Cousineau, D. (2005). Confidence intervals in within-subject designs: a simpler solution to loftus and masson's method. *Tutor. Quant. Methods Psychol.* 1, 4–45.

Cushman, F., and Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cogn. Sci.* 35, 1052–1075. doi: 10.1111/j.1551-6709.2010.01167.x

Deery, O., and Nahmias, E. (2017). Defeating manipulation arguments: interventionist causation and compatibilist sourcehood. *Philos. Stud.* 174, 1255–1276. doi: 10.1007/s11098-016-0754-8

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Q. J. Econ.* 75, 643–669.

Fitelson, B., and Hitchcock, C. (2011). "Probabilistic measures of causal strength," in *Causality in the Sciences*, eds P. M. Illari, F. Russo, and J. Williamson (Oxford: Oxford University Press), 600–627. doi: 10.1093/acprof:oso/9780199574131.003.0029

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. (2015). "How, whether, why: causal judgments as counterfactual contrasts," in *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Austin, TX.

Gerstenberg, T., and Lagnado, D. A. (2010). Spreading the blame: the allocation of responsibility amongst multiple agents. *Cognition* 115, 166–171. doi: 10.1016/j.cognition.2009.12.011

Gerstenberg, T., Lagnado, D. A., and Kareev, Y. (2010). "The dice are cast: the role of intended versus actual contributions in responsibility attribution," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* Austin, TX.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., and Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition* 177, 122–141. doi: 10.1016/j.cognition.2018.03.019

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009). Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition* 111, 364–371 doi: 10.1016/j.cognition.2009.02.001

Halpern, J. Y. (2016). *Actual causality*. Cambridge, MA: MIT Press.

Halpern, J. Y., and Hitchcock, C. (2015). Graded causation and defaults. *Br. J. Philos. Sci.* 66, 413–457. doi: 10.1093/bjps/axt050

Halpern, J. Y., and Kleiman-Weiner, M. (2018). "Towards formal definitions of blameworthiness, intention, and moral responsibility," in *Proceedings of the Thirty-Second Aaai Conference On Artificial Intelligence*, New York, NY, 1853–1860.

Hart, H. L. A., and Honoré, T. (1959). *Causation in the Law*. Oxford: Oxford University Press.

Hagmayer, Y., Meder, B., von Sydow, M., and Waldmann, M. R. (2011). Category transfer in sequential causal learning: the unbroken mechanism hypothesis. *Cogn. Sci.* 35, 842–873. doi: 10.1111/j.1551-6709.2011.01179.x

Hebbelmann, D., and von Sydow, M. (2017). Betting on transitivity in probabilistic causal chains. *Cogn. Process.* 18, 505–518.

Heider, F. (1958). *The Psychology Of Interpersonal Relations*. Hoboken, NJ: John Wiley and Sons Inc.

Hilton, D. J., McClure, J., and Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes. *Eur. J. Soc. Psychol.* 40, 383–400.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *Philos. Rev.* 116, 495–532. doi: 10.1215/00318108-2007-012

Hitchcock, C. (2012). Portable causal dependence: a tale of consilience. *Philos. Sci.* 79, 942–951. doi: 10.1086/667899

Hitchcock, C. (2017). "Probabilistic causation," in *The Oxford Handbook of Probability and Philosophy*, eds A. Hájek and C. Hitchcock (Oxford: Oxford University Press).

Hilton, D. J., and Slugoski, B. R. (1986). Knowledge-based causal attribution: the abnormal conditions focus model. *Psychol. Rev.* 93, 75–88. doi: 10.1037/0033-295x.93.1.75

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. LaSalle, IL: Reprinted by Open Court Press.

Icard, T. F., Kominsky J. F., and Knobe J. (2017). Normality and actual causal strength. *Cognition* 161, 80–93 doi: 10.1016/j.cognition.2017.01.010

Johnson, S. G. B., and Rips, L. J. (2015). Do the right thing: the assumption of optimality in lay decision theory and causal judgment. *Cognit. Psychol.* 77, 42–76. doi: 10.1016/j.cogpsych.2015.01.003

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., and Tenenbaum, J. B. (2015). "Inference of intention and permissibility in moral decision making," in *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Austin, TX.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., and Knobe, J. (2015). Causal superseding. *Cognition* 137, 196–209. doi: 10.1016/j.cognition.2015.01.013

Koskuba, K., Gerstenberg, T., Gordon, H., Lagnado, D. A., and Schlottmann, A. (2018). What's fair? How children assign reward to members of teams with differing causal structures. *Cognition* 177, 234–248. doi: 10.1016/j.cognition.2018.03.016

Lagnado, D. A., and Channon, S. (2008). Judgments of cause and blame: the effects of intentionality and foreseeability. *Cognition* 108, 754–770. doi: 10.1016/j.cognition.2008.06.009

Lagnado, D. A., Gerstenberg, T., and Zultan, R. (2013). Causal responsibility and counterfactuals. *Cogn. Sci.* 37, 1036–1073. doi: 10.1111/cogs.12054

Lagnado, D. A., and Gerstenberg, T. (2017). *Causation in Legal And Moral Reasoning inOxford Handbook of Causal Reasoning*. Oxford: Oxford University Press.

Lewis, D. (1973). Causation. *J. Philos.* 70, 556–567.

Lombrozo, T. (2010). Causal-explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions. *Cognit. Psychol.* 61, 303–332. doi: 10.1016/j.cogpsych.2010.05.002

Malle, B. F., Guglielmo, S., and Monroe, A. E. (2014). A theory of blame. *Psycholo. Inq.* 25, 147–186.

McClure, J., Hilton, D. J., and Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: probabilistic and social functionalist criteria for attributions. *Eur. J. Soc. Psychol.* 37, 879–901. doi: 10.1002/ejsp.394

Mikahil, J. (2007). Universal moral grammar: theory, evidence, and the future. *Trends Cogn. Sci.* 11, 143–152. doi: 10.1016/j.tics.2006.12.007

Moore, M. S. (2009). *Causation and Responsibility*. Oxford: Oxford University Press.

Murray, D., and Lombrozo, T. (2017). Effects of manipulation on attributions of causation, free will, and moral responsibility. *Cogn. Sci.* 41, 447–481. doi: 10.1111/cogs.12338

Nagel, T. (1979). *Moral Luck in Mortal Questions*. Cambridge: Cambridge University Press.

Pearl, J. (1999). Probabilities of causation: three counterfactual interpretations and their identification. *Synthese* 121, 93–149.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York, NY: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, reasoning and inference (2nd ed.)*. Cambridge, UK: Cambridge University Press.

Phillips, J., and Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proc. Natl. Acad. Sci. U.S.A.* 114, 4649–4654. doi: 10.1073/pnas.1619717114

Phillips, J., and Shaw, A. (2015). Manipulating morality: third-party intentions alter moral judgments by changing causal reasoning. *Cogn. Sci.* 39, 1320–1347. doi: 10.1111/cogs.12194

Faulkner, R. V. (1877). *Court of Crown Cases Reserved*. Washington, DC: American Chemical Society.

Royzman, E. B., and Baron, J. (2002). The preference for indirect harm. *Soc. Just. Res.* 15, 165–184.

Samland, J., and Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition* 156, 164–176. doi: 10.1016/j.cognition.2016.07.007

Sarin, A., Lagnado, D. A., and Burgess, P. W. (2017). The intention outcome asymmetry effect: how incongruent intentions and outcomes influence judgments of responsibility and causality. *Exp. Psychol.* 64, 124–141. doi: 10.1027/1618-3169/a000359

Shafer, G. (2000). Causality and responsibility. *Cardozo Law Rev.* 22:101123.

Sloman, S. A., and Lagnado, D. (2015). Causality in thought. *Annu. Rev. Psychol.* 66:223247. doi: 10.1146/annurev-psych-010814-015135

Spellman, B. A. (1997). Crediting causality. *J. Exp. Psychol. Gen.* 126, 323–348. doi: 10.1037/0096-3445.126.4.323

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

Sripada, C. S. (2012). What makes a manipulated agent unfree?. *Philos. Phenomenol. Res.* 85, 563–593. doi: 10.1111/j.1933-1592.2011.00527.x

Stephan, S., and Waldmann, M. R. (2018). Preemption in singular causation judgments: a computational model. *Top. Cogn. Sci.* 10, 242–257. doi: 10.1111/tops.12309

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub Co.

Sytsma, J., Livengood, J., and Rose, D. (2012). Two types of typicality: rethinking the role of statistical typicality in ordinary causal attributions. *Stud. Hist. Philos. Sci. Part C* 43, 814–820. doi: 10.1016/j.shpsc.2012.05.009

Vasilyeva, N., Blanchard, T., and Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cogn. Sci.* 42, 1265–1296. doi: 10.1111/cogs.12605

von Sydow, M., Hagmayer, Y., and Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Mem. Cogn.* 44, 469–487. doi: 10.3758/s13421-015-0568-5

Tadros, V. (2005). *Criminal Responsibility*. Oxford Oxford University Press.

Usher, M. (2018). Agency, teleological control and robust causation. *Philos. Phenomenol. Res.* 100:12537

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. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. (2006). Sensitive and insensitive causation. *Philos. Rev.* 115, 1–50. doi: 10.1215/00318108-2005-001

Zultan, R., Gerstenberg, T., and Lagnado, D. A. (2012). Finding fault: causality and counterfactuals in group attributions. *Cognition* 125, 429–440. doi: 10.1016/j.cognition.2012.07.014

Conflict of Interest: 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 June 2020
doi: 10.3389/fpsyg.2020.01054

Individuals vs. BARD: Experimental Evaluation of an Online System for Structured, Collaborative Bayesian Reasoning

Kevin B. Korb[1], *Erik P. Nyberg*[1], *Abraham Oshni Alvandi*[1], *Shreshth Thakur*[1], *Mehmet Ozmen*[2], *Yang Li*[1], *Ross Pearson*[1] *and Ann E. Nicholson*[1]*

[1] *Faculty of Information Technology, Monash University, Melbourne, VIC, Australia*, [2] *Department of Economics, University of Melbourne, Melbourne, VIC, Australia*

US intelligence analysts must weigh up relevant evidence to assess the probability of their conclusions, and express this reasoning clearly in written reports for decision-makers. Typically, they work alone with no special analytic tools, and sometimes succumb to common probabilistic and causal reasoning errors. So, the US government funded a major research program (CREATE) for four large academic teams to develop new structured, collaborative, software-based methods that might achieve better results. Our team's method (BARD) is the first to combine two key techniques: constructing causal Bayesian network models (BNs) to represent analyst knowledge, and small-group collaboration via the Delphi technique. BARD also incorporates compressed, high-quality online training allowing novices to use it, and checklist-inspired report templates with a rudimentary AI tool for generating text explanations from analysts' BNs. In two prior experiments, our team showed BARD's BN-building assists probabilistic reasoning when used by individuals, with a large effect (Glass' Δ 0.8) (Cruz et al., 2020), and even minimal Delphi-style interactions improve the BN structures individuals produce, with medium to very large effects (Glass' Δ 0.5–1.3) (Bolger et al., 2020). This experiment is the critical test of BARD as an integrated system and possible alternative to business-as-usual for intelligence analysis. Participants were asked to solve three probabilistic reasoning problems spread over 5 weeks, developed by our team to test both quantitative accuracy and susceptibility to tempting qualitative fallacies. Our 256 participants were randomly assigned to form 25 teams of 6–9 using BARD and 58 individuals using Google Suite and (if desired) the best pen-and-paper techniques. For each problem, BARD outperformed this control with very large to huge effects (Glass' Δ 1.4–2.2), greatly exceeding CREATE's initial target. We conclude that, for suitable problems, BARD already offers significant advantages over both business-as-usual and existing BN software. Our effect sizes also suggest BARD's BN-building and collaboration combined beneficially and cumulatively, although implementation differences decreased performances compared to Cruz et al. (2020), so interaction may have contributed. BARD has enormous potential for further development and testing of specific components and on more complex problems, and many potential applications beyond intelligence analysis.

Keywords: Bayesian networks, Delphi, CREATE, BARD, reasoning, decision-making, probability, uncertainty

OPEN ACCESS

Edited by:
York Hagmayer,
University of Göttingen, Germany

Reviewed by:
Joachim Funke,
Heidelberg University, Germany
Shenghua Luan,
Institute of Psychology (CAS), China
Leontien De Kwaadsteniet,
Radboud University Nijmegen,
Netherlands

****Correspondence:***
Ann E. Nicholson
ann.nicholson@monash.edu

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

Received: *08 September 2019*
Accepted: *27 April 2020*
Published: *18 June 2020*

Citation:
Korb KB, Nyberg EP, Oshni Alvandi A, Thakur S, Ozmen M, Li Y, Pearson R and Nicholson AE (2020) Individuals vs. BARD: Experimental Evaluation of an Online System for Structured, Collaborative Bayesian Reasoning. Front. Psychol. 11:1054. doi: 10.3389/fpsyg.2020.01054

1. INTRODUCTION

1.1. IARPA, CREATE, and BARD

Intelligence analysts are prone to the same reasoning mistakes as everyone else: groupthink, confirmation bias, overconfidence, etc. But when they produce bad assessments it can have disastrous results, such as the Weapons of Mass Destruction (WMD) reports used to justify the 2003 invasion of Iraq, which were later condemned by both sides of politics as analytically inadequate (United States Select Senate Committee on Intelligence, 2004; Silberman and Robb, 2005). So, the US intelligence community's research body, IARPA (Intelligence Advanced Research Projects Activity)[1] has sought "structured analytic techniques" that would methodically produce better reasoned intelligence reports. Their latest, multi-million-dollar program was CREATE (CRowdsourcing Evidence, Argumentation, Thinking, and Evaluation)[2], which specifically sought *software-based* approaches to enable *crowdsourced* structured techniques, and funded four large academic teams to pursue contrasting approaches to this end.

Our BARD team (Bayesian ARgumentation via Delphi)[3] included computer scientists at Monash (led by Kevin Korb, Ann Nicholson, Erik Nyberg, and Ingrid Zukerman) and psychologists at UCL and Birkbeck (led by David Lagnado and Ulrike Hahn) who are experts in encoding people's knowledge of the world in maps of probabilistic causal influence: causal Bayesian Networks (BNs). A good map can provide the logical skeleton of a good intelligence report, including the probabilities of competing hypotheses, the impact of supporting evidence, relevant lines of argument, and key uncertainties. Two well-known difficulties here are eliciting sufficient analyst knowledge and amalgamating diverse opinions. So, our team also included psychologists from Strathclyde (led by Fergus Bolger, Gene Rowe, and George Wright) who are experts in the Delphi method, in which a facilitator methodically leads an anonymous group discussion toward a reasoned consensus.

The outcome of our research is the BARD system: an application and methodology whose two defining features are the construction of causal BNs and a Delphi-style collaborative process, with the aim of producing better reasoning under uncertainty and expressing it clearly in written reports. In addition, we incorporated several other features likely to improve performance, most notably: an anytime audiovisual training package, a guided incremental and iterative workflow, report templates to encourage analysts to include items often neglected, and the auto-generation of natural language text expressing some of the BN's key features. We provide a brief sketch of the system in section 3; for a more detailed picture see Nicholson et al. (2020)[4].

1.2. CREATE Experiments on BARD

A key feature of IARPA's approach is the use of external testing, so their independent testing team designed a major experiment to test the effectiveness of the four CREATE approaches, including BARD. We developed, tested, and contributed some new reasoning problems that captured key elements of intelligence analysis in a simpler form, which were reviewed and included in the IARPA suite of test problems. IARPA deemed the appropriate control condition to be individuals using the Google Office Suite, since this mirrored "business as usual" for intelligence analysis. Unfortunately, IARPA's testing team relied upon retaining a large number of volunteer participants who were not significantly compensated, and attrition was so high (regardless of which of the four systems participants used) that the experiment was terminated early without obtaining any statistically useful data.

Anticipating this outcome, we designed and carried out the present study, relying on a smaller number of participants who received significant compensation. To date, it constitutes the only significant and critical experimental test of the entire BARD system used end-to-end on reasoning problems developed for CREATE. The study methodology is described in section 4, with results and discussion presented in sections 5 and 6.

Since BARD is multifaceted, and our small study is necessarily limited in the variables manipulated, it does not show how much each facet contributed to the total result. None of them are statistical confounds for this experiment, since the aim always was to test BARD as a whole. However, the contribution of each facet—and how to polish them further so they shine better together—are further research questions of great interest. In section 2, we briefly review the most relevant theory and previous experimental results, including two experiments our team performed to separate BARD's BN construction from its Delphi collaboration. This review supports the view that each of BARD's facets most likely contributes *positively and cumulatively* to total BARD performance. We hope that future research will improve, validate, and measure each contribution.

2. BACKGROUND

2.1. Intelligence Analysis Problems

Intelligence analysis typically requires assessing the probability of some conclusion based on available pieces of evidence, and writing reports for decision-makers to explain that assessment. To express those probabilities, US analysts are expected to use a standard verbal terminology corresponding to defined numerical ranges (e.g., "very likely" means 80–95%) as specified in ICD-203, which "establishes the Intelligence Community (IC) Analytic Standards that govern the production and evaluation of analytic products" [Office of the Director of National Intelligence (ODNI), 2015]. The same conclusions are often reassessed periodically as new evidence arises. This sort of intelligence analysis requires a type of reasoning under uncertainty that is not unusual: similar reasoning is required in many other domains, and we hope that BARD's success with our test problems will ultimately be transferable to many real-world problems.

To test the BARD system, our team needed to develop new reasoning problems that captured the key elements of intelligence

[1] IARPA is an organization within the Office of the Director of National Intelligence responsible for leading research to overcome difficult challenges relevant to the United States Intelligence Community. For more details see https://www.iarpa.gov/.

[2] CREATE webpage: https://www.iarpa.gov/index.php/research-programs/create.

[3] BARD webpage: http://bard.monash.edu/.

[4] To view and interact with BARD (including solved demonstration problems), to view the BARD training e-course, or to use BARD for academic research purposes (including refereeing and replications), please email Prof. Ann Nicholson at ann.nicholson@monash.edu.

analysis in a simpler form. Basic scenarios and evidence are presented in written form, and answers must ultimately be given in written form, but participants can use other means (e.g., BNs, pen-and-paper calculations) in between. In each of our short reasoning problems, we incorporated a major reasoning difficulty likely to lead to some qualitatively incorrect conclusions and explanations, and we also tested the accuracy of quantitative estimates. Our reasoning problems were developed and tested by our London-based cognitive psychologists.

We used two of these problems in this experiment. The Kernel Error problem involves the cognitive difficulty known as "explaining away." For example, if my wet lawn must be caused by either a sprinkler or rain (or both), and these two causes are each sufficient and otherwise independent, then seeing the wet lawn raises the probability of both possible causes. However, if I discover that it rained, this entirely "explains away" the wet lawn, and the probability of the sprinkler should be lowered to its initial value. Our team's psychology experiments with Kernel Error formally confirmed what computer scientists have informally observed: people have difficulty readjusting their probabilities appropriately (Liefgreen et al., 2018).

The Cyberattack problem involves the cognitive difficulty known as dependent evidence. For example, how much additional weight should we give to a second medical test result if we know that the second test was of the same type as the first? This depends on how the results are correlated, e.g., how often errors in the first test will be caused by factors that will also cause errors in the second test. Even if people have precise figures for this, our team's psychology experiments with Cyberattack formally confirmed that people find it difficult to combine dependent evidence accurately (Pilditch et al., 2018).

2.2. Probabilistic and Causal Reasoning Errors

Psychological research has revealed many difficulties people have with both probabilistic and causal reasoning (Kahneman et al., 1982; Hahn and Harris, 2014; Newell et al., 2015). To summarize a very large literature:

- One general factor that increases the probability of such errors is simply complexity. Facing a mass of interconnecting evidence and long lines of argument, it is easier to make an error somewhere along the line in assessing the impact of evidence on a conclusion.
- Another general factor is specific dependence patterns that people find surprisingly difficult. Besides explaining away and dependent evidence, these include "screening off," i.e., when knowledge of the state of a common cause renders two dependent effects independent of each other, and mistaking correlation for direct causation when a hidden common cause is far more likely (Gopnik et al., 2001; Lagnado and Sloman, 2004; Kushnir et al., 2010; Pearl and Mackenzie, 2018).
- A third general factor is the common biases in the way people express and update their probabilities, such as overconfidence, i.e., exaggerating the probability of likely events and the improbability of unlikely events (Moore and Healy, 2008); conservative updating, i.e., inadequately weighting new evidence when revising beliefs (Kahneman et al., 1982; Matsumori et al., 2018); base-rate neglect, i.e., inadequately weighting the priors (Welsh and Navarro, 2012); and anchoring, i.e., depending too much on an initial piece of information (the anchor) (Kahneman et al., 1982).

2.3. BNs to Reduce Reasoning Errors

A key reason for IARPA's interest in structured representations is to reduce such cognitive difficulties when analyzing problems (Heuer, 1999). Causal BNs are particularly well-suited for the task, since they explicitly represent and accurately combine both probabilistic and causal information.

Formally, a BN is a directed, acyclic graph whose nodes represent random variables, and whose arrows represent direct probabilistic dependencies, often quantified by conditional probability tables (CPTs) associated with each node. In causal BNs, each of these arrows also represents direct causal influence—hence, they can also predict the effects of decisions to intervene. Users can enter exact or uncertain evidence about any variables, which is then efficiently propagated, updating the probability distributions for all variables. Thus, causal BNs can support and perform predictive, diagnostic (retrodictive), explanatory, and decision-oriented probabilistic reasoning. For more technical details, see Pearl (1998), Spirtes et al. (2000), and Korb and Nicholson (2011).

But how does constructing a BN help people avoid reasoning errors, rather than merely reproducing them? Reasoning errors aren't bad beliefs; they are bad ways to develop or combine beliefs. So, BN assistance doesn't depend on all the analysts' beliefs being true, it just enables analysts to accurately draw the conclusions that are implied by their own beliefs. It's analogous to using a calculator to help avoid arithmetical errors: provided that people enter the numbers and operations they believe are correct, the calculator can be relied upon to combine them accurately. In constructing BNs, analysts must explicitly think about and identify the causal structure (rather than make implicit assumptions about it). The model then requires all the relevant probabilities to be entered (so none of these can be neglected). The BN calculations then automatically avoid almost all the errors discussed above. The way modeling with BNs helps avoid errors has been explained and/or empirically verified for multiple specific reasoning difficulties: base-rate neglect (Korb and Nyberg, 2016); confusion of the inverse, i.e., interpreting the likelihood as a posterior (Villejoubert and Mandel, 2002); the conjunction fallacy, i.e., assigning a lower probability to a more general outcome than to one of the specific outcomes it includes (Jarvstad and Hahn, 2011); the jury observation fallacy, i.e., automatically losing confidence in a "not guilty" verdict when a previous similar conviction by the defendant is revealed (Fenton and Neil, 2000); and most recently, the zero-sum fallacy, i.e., not recognizing when a piece of evidence increases the probability of *both* a hypothesis and its most salient rival (Pilditch et al., 2019). Exceptions to this rule might be reasoning errors that arise from mistaken ways to express individual beliefs, e.g., ambiguities in variable definitions, or overconfidence in the initial probabilities assigned. For such issues, the critical discussion engendered by structured social processes may be more useful, per sections 2.4 and 2.5.

More generally, given their ability to embody normatively correct reasoning, causal BNs have been used to analyze common fallacies in informal logic (Korb, 2004), analyze and assess a variety of arguments in criminal law—where they have exposed some common errors in evidential reasoning (e.g., Fenton et al., 2013; Lagnado et al., 2013), analyze human difficulties with reasoning under uncertainty (e.g., Hahn and Oaksford, 2006; Hahn, 2014), model human knowledge acquisition while solving complex problems (e.g., Holt and Osman, 2017), and as a proposed method for argument analysis (Korb and Nyberg, 2016). In practical contexts, they have been deployed to support human reasoning and decision making under uncertainty in such diverse domains as medicine (e.g., Flores et al., 2011; Sesen et al., 2013), education (e.g., Stacey et al., 2003), engineering (e.g., Choi et al., 2007; Bayraktar and Hastak, 2009; Misirli and Bener, 2014), surveillance (e.g., Mascaro et al., 2014), the law (e.g., Fenton et al., 2013; Lagnado and Gerstenberg, 2017), and the environment (e.g., Chee et al., 2016; Ropero et al., 2018).

Many BN software tools have been developed to assist in building, editing, evaluating, and deploying BNs. These include Hugin[5], GeNie[6], Netica,[7] AgenaRisk[8], BayesiaLab[9], and a plethora of research software tools, e.g., Elvira[10], R BN libraries[11], BNT[12], SamIam[13], and BayesPy[14]. However, all of these tools assume that the user understands BN technology (or they offer only rudimentary help), and they assume the user knows how to translate their knowledge of a causal process or argument into a Bayesian network. In the BARD system, we improved on this first generation of BN tools by providing far better training and guidance (see section 3.2), and by providing a structured workflow that draws on new BN "knowledge engineering" concepts and best practices (see section 3.3).

2.4. Delphi Groups to Improve Reasoning

There is considerable evidence that decision making by groups, either by reaching consensus or by amalgamation, can produce better outcomes than decision making by individuals (e.g., Salerno et al., 2017; Kugler et al., 2012; Charness and Sutter, 2012; Straus et al., 2011). However, there are also well-known problems that arise with group interactions, e.g., anchoring, groupthink, and psycho-social influences (for more details, see Kahneman et al., 1982; Mumford et al., 2006; Packer, 2009; Stettinger et al., 2015). Groups also have potential logistical advantages in that subtasks can be divided among members and/or performed by the most competent.

A number of methods have been developed over the years that attempt to harness the positives of groups while preempting or ameliorating the negatives. One of the best-known is the Delphi technique (e.g., Linstone and Turoff, 1975), an example of a "nominal group" technique: the group members never actually meet, but rather, interact "remotely." The defining characteristics of a Delphi process (e.g., Rowe et al., 1991) are: *anonymity* to reduce the influence that powerful or dogmatic individuals can have on group judgments; *iteration* with *feedback*, which allows participants the chance to reconsider and improve their responses in the light of information from other group members; and *aggregation* (or collation, if responses are qualitative in nature) of group responses, often done by a *facilitator*—who can also assist by reducing unproductive exchanges and encouraging task completion (but avoids making original contributions). At least for short-term forecasting problems and tasks involving judgements of quantities, Delphi has generally shown improved performance compared to freely interacting groups or a statistically aggregated response based on the first-round responses of individual participants (Rowe et al., 1991).

2.5. Delphi for Constructing BNs

Recently, one study used a form of Delphi for point-estimate CPT elicitation (Etminani et al., 2013), while for BN structure elicitation Serwylo (2015) pioneered online crowdsourcing and automated aggregation (albeit non-Delphi). Some of the present authors proposed a Delphi-style elicitation of BN structure in an epidemiological case study (Nicholson et al., 2016). However, BARD is the first system to use Delphi for developing and exploring an entire BN model, including variables, structure and parameters, and also for more complex reasoning problems.

The major difficulty in using Delphi here is that both the workflow and the output are complex: the workflow necessarily involves multiple, logically dependent steps, and users should be encouraged to improve their complex answers iteratively by repeating steps. One approach would be for each participant to complete the entire process before discussing their work with others, but this means they would learn nothing from others during the process and have complicated outputs to assess and discuss at the end. Another approach would be to use a traditional Delphi process at each step and make the workflow strictly linear, but this loses all the advantages of iterative development, and requires synchronized participation. BARD resolved this dilemma by using a compromise: "Real-Time Delphi" (see section 3.3). One crucial achievement of this experiment is to demonstrate the feasibility of combining Delphi with BN construction in this way.

In real-world applications, the relevant probabilities may come from either data, such as available studies on the false positive and false negative rates for a medical test, or expert opinion, such as the relative risks of new medical treatments where there is little data available. In either case, there may be disagreement and uncertainty. Instead of a single point probability, the available information is then better summarized as some sort of probability distribution or interval, which may be interpreted as meta-uncertainty about the appropriate point

[5] Hugin website: https://www.hugin.com/.

[6] GeNie website: https://www.bayesfusion.com/.

[7] Netica website: https://www.norsys.com/index.html.

[8] Agena Risk website: https://www.agenarisk.com/.

[9] BayesiaLab website: http://www.bayesia.com/.

[10] Elvira website: http://leo.ugr.es/elvira/.

[11] R BN website: http://www.bnlearn.com/.

[12] BNT website: https://github.com/bayesnet/bnt.

[13] SamIam website: http://reasoning.cs.ucla.edu/samiam/.

[14] BayesPy website: https://pypi.org/project/bayespy/.

probability, and called a "vague" probability. There have been various protocols proposed for eliciting and combining such probabilities from multiple experts, such as 3-point methods (e.g., Malcolm et al., 1959; Soll and Klayman, 2004), a 4-point method (Speirs-Bridge et al., 2010), and the IDEA protocol (Hemming et al., 2018a); however, these have not been integrated into any of the commercial or research BN software tools. Instead, these protocols are applied externally to the BN software and then incorporated by the BN model builder (e.g., Nicholson et al., 2011; van der Gaag et al., 2012; Pollino et al., 2007; Hemming et al., 2018b). Uniquely, BARD integrates elicitation tools of this kind with BN construction. However, the test problems in this experiment specify appropriate point probabilities in the problem statements, in order to simplify the task and yield uncontroversial, normatively correct solutions. So, assessing the effectiveness of BARD for vague probabilities must await future research.

2.6. Checklists for Improving Reasoning

One of the simplest structured techniques is the checklist—yet, it has proven highly effective in reducing errors in such challenging expert tasks as piloting aircraft and, more recently, in performing medical surgery (Russ et al., 2013). Effective checklists are carefully designed to provide timely and concise reminders of those important items that are most often forgotten. For CREATE, IARPA had already identified important general elements of good reasoning that are frequently omitted, e.g., articulating competing hypotheses, and noting key assumptions (Intelligence Advanced Research Projects Activity, 2016). This suggested that something like a reasoning checklist could be useful, if added to the BARD system.

For BN-building, the functions of a checklist are implicitly fulfilled by our stepwise workflow with step-specific tips, and associated automated reminders. For report-writing, we implemented the checklist idea more explicitly in the form of a report template with section-specific tips, and associated automatic text generation (see section 3.4).

2.7. Experiments Separating BN Construction From Delphi Collaboration

The BARD team performed two other critical experiments on the BARD system, reported in detail in the references below, which provide some evidence that its two principal features—BN construction and Delphi groups—both contribute positively and cumulatively to BARD's total performance.

In the SoloBARD experiment (Cruz et al., 2020), individual participants used a version of BARD without any social interaction to solve three of the reasoning problems our team developed, including the Kernel Error and Cyberattack problems used in this experiment. The control condition consisted of individuals provided only with Microsoft Word and IARPA's generic critical thinking advice. The results showed much better performance from the individuals using SoloBARD. This provides some evidence for the feasibility and effectiveness of BN construction (supported by BARD's other non-social features, such as templates) to analyze probabilistic reasoning problems and produce written reports.

In the Structure Delphi experiment (Bolger et al., 2020), individual participants who had previously used BARD were asked to analyze some of our other reasoning problems they had not previously seen, but only for the critical and most distinctive subtask in BN construction: selecting the right variables and causal structure. All other subtasks in solving our reasoning problems are similar to tasks for which Delphi has already been shown to be effective in prior literature. Individual participants were shown the structures purportedly proposed by other members of their Delphi group (although, in fact, generated earlier by similar participants and curated prior to the experiment) and invited to rate these structures and revise their own. The results showed they made substantial improvements over their initial responses, both in the top-rated structures and in the revised structures. This provides some evidence that BARD's Real-Time Delphi social process is an additional positive contributor to performance in analyzing probabilistic problems.

We did not perform any experiment directly comparing groups using BARD to groups using Google Docs. This was partly due to our resource limitations, and also to its lower prioritization by IARPA. Given our combined experimental results, we think it highly unlikely that groups with Google Docs could have outperformed groups with BARD on these particular probabilistic reasoning problems. Nevertheless, sorting out the exact independent and combined contributions of BARD's BN-building and structured social processes vs. unaided, unstructured group processes remains an interesting research task for the future.

3. THE BARD APPLICATION

3.1. Overview

BARD (Nicholson et al., 2020) supports the collaborative construction and validation of BN-based analyses in a web application, in a Delphi-style workflow. Analysts in small groups, optionally assisted by a facilitator, are guided through a structured Delphi-like elicitation protocol to consider and represent their relevant knowledge in a causal BN augmented by descriptive annotations. BARD provides tools to assist the elicitation of a causal BN structure and its parameters, review and build consensus within the group and explore the BN's reasoning in specific scenarios. BARD encourages analysts to incrementally and iteratively build their individual BNs and seek regular feedback through communication with other group members and the facilitator. The group may decide to adopt the highest-rated individual BN or a facilitator can assist in the production of a consensus model. From an individual or group BN, BARD auto-generates an outline of a structured verbal report explaining the analysis and identifying key factors (including the diagnosticity of evidence and critical uncertainties). Analysts and the facilitator can revise this into an intuitive narrative explanation of the solution, using a structured template prompting users to incorporate elements of good reasoning.

3.2. Better Training for Building BNs

Our experience is that substantial training is required to model effectively with BNs. For example, the standard BayesiaLab training is conducted in a 3-day course[15], while Bayesian Intelligence Pty Ltd offers 2-day training as standard[16]. However, the requirements of testing and evaluation in the IARPA CREATE program limited upfront training to 4 h of online, individual, self-paced training, without any input or assistance from a human instructor.

BARD upfront training developed for CREATE is delivered as condensed but high-quality audiovisual e-courses, with corresponding practical exercises, example solutions, and context-sensitive help and tips embedded in the software. They cover the fundamentals of Bayesian network modeling, how teams function in BARD, the differences in the responsibilities of facilitators and analysts, and details on how to use the BARD software itself[17].

3.3. Better Workflow for Building BNs

The BARD workflow decomposes the task into a logical series of six smaller steps (see **Figure 1**). Step 1 focuses on understanding the problem for analysis, particularly identifying hypotheses as well as the most relevant factors and evidence. In Steps 2 to 5, participants build a BN model of the analytic problem, broken down into variable selection (Step 2), adding arrows to define the structure (Step 3), parameterizing the model to specify the probabilities (Step 4), and then exploring and validating the BN's reasoning on specific scenarios (Step 5). Finally, the participants individually and collectively construct a written report (Step 6).

At each of the steps, analysts are required to first work on their solution in isolation (blinded to other responses) and then "Publish" their work (which makes it available to other analysts), before they can view other analysts' work or the current group solution (produced by the facilitator), and discuss them via the step-specific discussion forum (see **Figure 2**). Publishing also allows analysts to move forward to the next step. When present, a facilitator's role is to: support the team's progression through the steps in terms of timeliness and focus; optionally synthesize the team's work in the "group" solution (both BN and report) with minimal original contributions of their own; encourage review, feedback, and discussion; and submit the final analytic report.

Thus, apart from the initial response requirement, at each step group members are free to progress to subsequent steps at their own pace and can move flexibly backwards and forwards between steps. This BARD workflow is based on a "roundless" Delphi variant called "Real-Time Delphi" (Gordon and Pease, 2006), where the sub-steps of providing individual responses, viewing information from other participants, and improving responses are not controlled by the facilitator, but rather, where the transitions occur immediately, i.e., in "real time." This allows far more flexibility about when the participants can make their contributions and speeds up the Delphi process, since analysts do not have to wait for the facilitator to amalgamate or collate responses, as well as reducing the need for facilitation. It also allows users to return to earlier steps to expand on their answers, since BNs are best built iteratively and incrementally (Laskey and Mahoney, 1997, 2000; Boneh, 2010; Korb and Nyberg, 2016). The trade-off is that, since the participants can see each other's responses directly, rather than after amalgamation or collation, some of the biases deriving from direct interaction that Delphi is designed to eliminate may re-emerge.

At Step 6, analysts can also rate their own and other analysts' reports on a 10-point scale; after rating a report, they can see their own rating and the current average rating. This feature was introduced as a quantitative high-level assessment to help focus discussion, as well as providing guidance to the facilitator on which report(s) to use as the basis for the team solution. However, in the absence of a facilitator, these ratings can also be used as input to an algorithm to automatically select an individual report as the team solution (see section 4.3.2).

Using this workflow, a team can methodically produce an analytic report explaining the members' collective answer to the problem and their reasoning behind it.

3.4. Report Templates and Automated BN Explanations

BARD pre-populates the written report workspace with a few generic headings, along with explanatory tips for each heading. These function as checklist-style reminders and placeholders for these general elements, e.g., the relevant hypotheses and their prior probabilities, and they also clarify the presentation for the reader. Participants are encouraged to include tables or figures, such as an image of the BN structure, if these enhance clarity further. We note that TRACE, one of the other four CREATE projects, also experimented successfully with flexible report templates (Stromer-Galley et al., 2018), which supports the view that they make some positive contribution.

In conjunction, we developed a rudimentary AI tool for generating text explanations of the relevant BN features, and organized this text under the same template headings so that it could readily be copied or imitated in the written reports. The reason for providing such assistance is that, especially when BNs become more complex, it can be difficult to understand the interaction between evidence items and their ultimate impacts on the conclusion. Although the BN will have calculated this accurately, well-reasoned reports demand that the impact be explained verbally, and it helps if the BN can explain itself[18].

4. METHODOLOGY

This study was approved by the Monash University Human Research Ethics Committee, with the plan[19] lodged with the

[15] BayesiaLab website: http://www.bayesia.com/events.

[16] Bayesian Intelligence website: https://bayesian-intelligence.com/training/.

[17] For a glimpse of this upfront training, see "BARD Screenshots" at https://tinyurl.com/bard-publications.

[18] This "explainable AI" (XAI) feature is now undergoing further development as part of a spinoff project involving several BARD researchers. "Improving human reasoning with causal Bayes networks: a multimodal approach"is a major 3-year project at Monash University and the University of London funded by the Australian Research Council. See https://dataportal.arc.gov.au/NCGP/Web/Grant/Grant/DP200100040.

[19] "Experiment Design" available at: https://bit.ly/2OBVBCc.

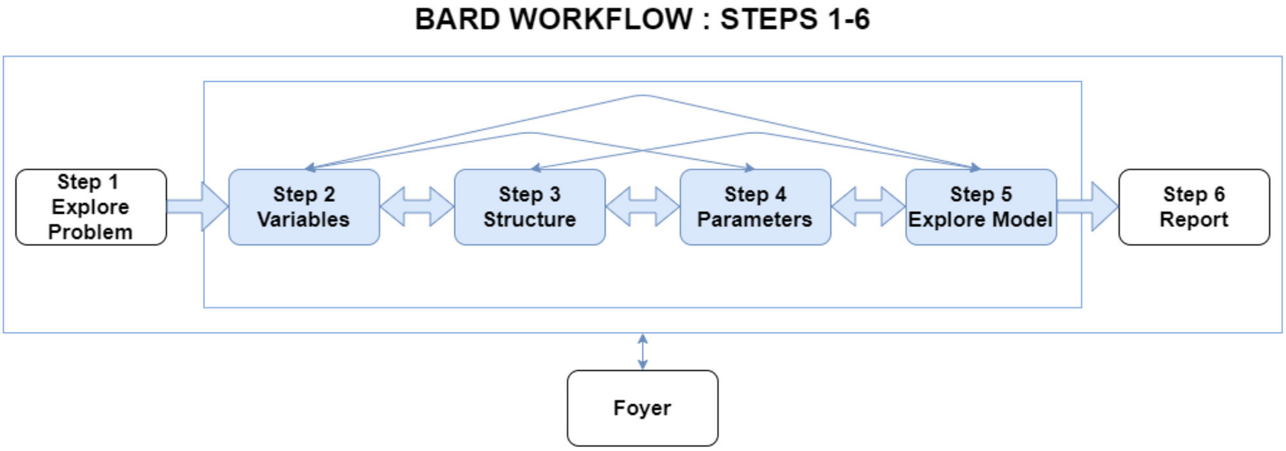


FIGURE 1 | BARD's workflow consists of six steps. Analysts and the facilitator can move flexibly backwards and forwards between steps.

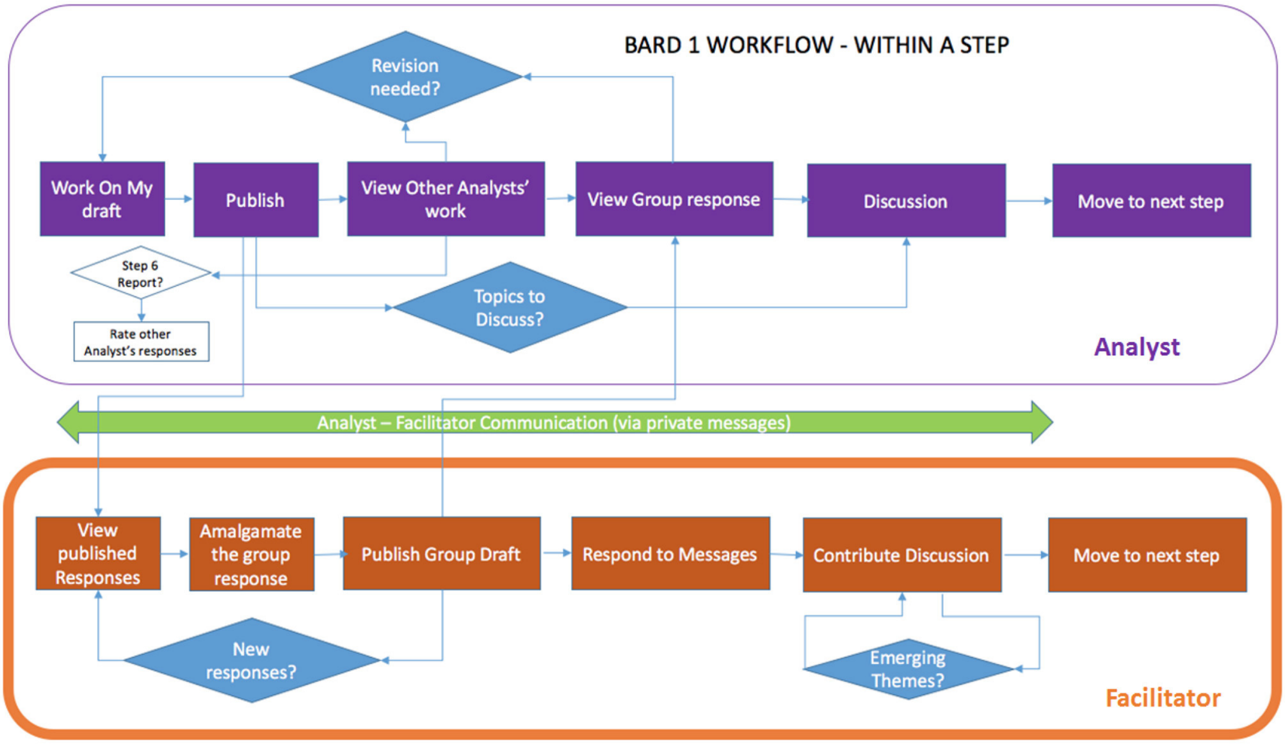


FIGURE 2 | High level representation of the BARD workflow within a step for analysts and the facilitator.

Open Science Framework (OSF)[20] and approved by them and by the IARPA CREATE program.

4.1. Participants

Power Analyses: We conducted separate power analyses for our t-tests and repeated measures ANOVAs, both assessed for a statistical power of 0.8. We chose large effect sizes to reflect that only substantial improvements over the control would be sufficient to justify adopting the BARD system. For the t-test of the difference between two independent means at the 5% level of significance (one-sided), with equal sample sizes and a very large effect size (Cohen's $d = 1.33$, equivalent to a 20% improvement of BARD over a control mean score of $\mu_k = 20$ with equal $\sigma = 3$ across conditions), we calculated an indicative sample size of 16 score observations split evenly between BARD and the control.

[20] The Open Science Framework is an open source software project that facilitates open collaboration in science research. See for more information: https://osf.io/.

Assuming 8 individuals were recruited into each BARD team to form a single test score, this would require 72 individuals in total. Assuming a larger standard deviation ($\sigma = 5$), a large effect size ($d = 0.8$) and retaining other assumptions, we calculated 21 observations per condition requiring 189 individuals in total. For a repeated measures within-factors ANOVA at the 5% level of significance (one-sided), with equal sample sizes, across two periods, with two factors and a very large effect size (partial $\eta^2 = 0.735$, equivalent to Cohen's $d = 1.33$), we calculated an indicative sample size of 4 observations per condition requiring 36 individuals in total. For a large effect size (partial $\eta^2 = 0.39$, equivalent to Cohen's $d = 0.8$) we calculated 8 observations per condition or 72 individuals in total. To cater for the worst case among these analyses, we set a minimum recruitment target of 189 individuals.

Recruitment Methods: We recruited participants via social media (using Monash University's Facebook, LinkedIn, and Twitter pages to advertise for volunteers), Monash University student organizations, and the Monash Psychology department's SONA system. All participation was voluntary, and participants could withdraw at any time prior to completion of the study while retaining any compensation earned. All responses were fully anonymized (including all IP address information). A short quiz[21] including probabilistic reasoning questions and personality questions was completed by all participants when registering for the experiment. This was not used to select applicants for the experiment or assign their condition, but was used later to help allocate the facilitator role within BARD teams.

Sample Size, Conditions, and Demographics: We attempted to over-recruit since the rate of attrition could not be known in advance, and we succeeded in obtaining 295 registrations. These potential participants were 18–57 years old with a mean of 29.7 and a standard deviation of 7.2 years. By optional self-identification, there were 139 females, 141 males, and 15 others. The target population was English-speaking adults with some undergraduate experience, so individuals who had not yet completed high school education or were younger than 18 were excluded.

Following the randomized control trial (RCT) standard, these potential participants were selected randomly into two conditions for between-condition comparisons. After asking them to confirm their availability for the respective time commitments required, we began the experiment with 256 participants. 58 control (K) participants were asked to work individually at any time to produce reports, using the Google Suite tools and (if desired) some pen-and-paper techniques (see section 4.3.3). The remaining 198 experimental (X) participants were asked to work collaboratively and synchronously in teams of 6–9 using the BARD tool. By self-identification, K contained 20 females, 37 males, and 1 other, while X contained 98 females, 86 males, and 14 others.

Participants were kept blind of their condition in the sense that they were not informed about the nature of any other conditions. However, blinding was necessarily imperfect, in that many participants would have heard of the IARPA CREATE program and/or BARD independently of the experiment itself, and may have been aware that the BARD project utilizes BN technology. In particular, some K participants may have been aware that they were not using the technology under development and performing as controls. Of course, every participant was trained explicitly only in the tools actually required for their condition.

Compensation: Participants were compensated for adequate participation in each session in the form of a GiftPay[22] voucher, and those who participated in all sessions received a bonus. All participants (X and K) were required to complete the upfront training to receive compensation. In each of the 5 problem-solving weeks, X participants were required to attend joint problem solving sessions and actively work on their reports to receive compensation, while K participants were only required to complete their report. For the optional webinars (see section 4.3), attendance was sufficient.

4.2. Materials

Three analytic problems were selected for the study; all were probabilistic in nature and ideally suited to being solved using Bayesian networks: (A) Smoking and Cancer[23]; (B) Kernel Error (Liefgreen et al., 2018); and (C) Cyberattack (Pilditch et al., 2018).

All problems had corresponding marking rubrics, with those for B and C developed previously by our team's cognitive psychologists, and a similar format used here for A[24]. Participants were explicitly asked to provide some specified probabilities, but also asked to justify those answers. In the rubrics, assessors were provided with both the correct answers and a short list of specific observations which ought to feature in any sound and thorough justification, e.g., that one evidence source is more reliable than another. Assessors awarded one point for each answer and each observation that participants fully included, and a half point for each observation that was only partially included. The final rubric score was simply the sum of these points.

The nature of these rubrics entails that there are a large number of available points (13, 38, and 34 respectively), but the mean proportion of these points obtained by participants tends to be low. It is less clear to participants which items to include in their justifications than in their answers, and in ordinary life people frequently give shorter, partial justifications that leave some relevant facts unstated. Hence, even participants with correct answers obtained by correct reasoning are likely to omit some point-scoring observations from their justifications. Conversely, even participants who give incorrect answers or use incorrect reasoning are likely to score some points in their justifications. To avoid this "random noise" inherent to scoring justifications, the SoloBARD experiment (Cruz et al., 2020) also compared points scored only from the answers to the explicit questions, and found a much greater effect size (Glass' $\Delta = 1.4$) in favor of SoloBARD—but we did not propose or perform this analysis for our BARD experiment.

Unlike the SoloBARD experiment, we split Problems B and C into two parts. Part 1 introduced a new scenario with relevant

[21] "Short Quiz" available at: https://bit.ly/2K6Nj6N.

[22] GiftPay website: https://www.giftpay.com/.

[23] "Smoking and Cancer - Problem Statement" available at: https://bit.ly/2V2rwl4.

[24] "Smoking and Cancer - Rubric" available at: https://bit.ly/2v92emU.

TABLE 1 | Experiment schedule by week: for webinars, training, and problem-solving.

Week	Webinar	Task
0	Welcome	Training
1	Q&A on Training	(A) Smoking and Cancer
2	Solution and Q&A for (A)	(B) Kernel Error, Part 1
3	–	(B) Kernel Error, Part 2
4	Solution and Q&A for (B)	(C) Cyberattack, Part 1
5	–	(C) Cyberattack, Part 2

evidence and questions that needed to be answered. Part 2 of each problem was presented in the following week, building on the first by adding new evidence to the problem descriptions and then asking additional questions about its impact. BN models readily allow for such "phased" problems, and BARD takes advantage of that in allowing "scenarios" to be built incrementally along with the models used to analyze them. So, both K individuals and X teams were able to build on their analyses for Part 1, even though those questions were not repeated and their rubric scores did not carry over to Part 2.

In Part 2 of both problems, participants must cope with more variables and more dependencies between them—which makes the problems computationally more difficult than in Part 1. Furthermore, these additional elements introduce the major cognitive difficulties designed into these problems. For both reasons, the Part 2 questions should be more difficult for participants, and we expected them to achieve a lower proportion of the available marks. Furthermore, we expected the advantage of using BARD to become more pronounced. To test this secondary hypothesis, we used a separate ANOVA for each of these problems to detect any significant interaction, despite the small loss of statistical power in detecting the main effect.

In the SoloBARD experiment, Problems B and C (not divided into parts) seemed to present roughly the same difficulty for participants: controls obtained roughly the same proportion of the available points in both problems, and so did participants using SoloBARD. Our Problem A was structurally comparable to the first part of the other two problems, and hence not particularly difficult nor divided into parts. It is similar to example BN problems common in introductory undergraduate Artificial Intelligence courses, and partly intended to provide additional training for both X and K in conjunction with the associated webinar on how it can be accurately solved, before they proceeded to the more difficult problems.

The problem-solving was conducted over 5 consecutive weeks, with the webinars, training, and problems being presented in the sequence shown in **Table 1**.

4.3. Design and Procedure

4.3.1. The Variables

The variable under manipulation was the tool and associated training used for analyzing problems and writing solutions; the dependent variable assessed was performance in producing these solutions. X and K membership was assigned uniformly randomly, using random.org to select a sufficient number of participants for K. This implicitly controlled for other independent variables; those measured, via the registration quiz and BARD's usage monitoring, were: Education level (high school, some college, BA, MS, PhD); Probability/Stats education; Sex; Nationality; Age ($\geq$ 18); Total login time.

Very high attrition rates were observed in all preliminary studies by CREATE teams, including pilot studies for this experiment in both X and K: up to 50% per week, which would have been unsustainable over the course of the experiment. We made several adjustments to minimize and cater for attrition, most notably by encouraging frequent social engagement. X team members were required to work synchronously; and for both X and K we introduced "webinars" (i.e., online seminars) presented by a member of the experimental team that provided additional training and Q&A; these were voluntary, but (apart from the initial Welcome) participants received additional compensation for attendance.

4.3.2. Experimental Condition (X)

Training: For this study, compulsory upfront X training consisted of only 2 h of the BARD e-courses for analysts, delivered individually using a Learning Management System (Moodle). Participants were then asked if they were willing to take on the facilitator role. Those who answered "yes" and completed the short, optional facilitator e-course were subsequently considered as prospective facilitators. All e-courses remained accessible via the BARD platform throughout the experiment.

The four different webinars were held according to the week-by-week schedule in **Table 1**, and within each week, the scheduled X webinar was presented four times on weekday evenings to cater for participant availability and keep the numbers in each session manageable. Their respective aims were: to welcome and introduce participants to the experiment and encourage them to do the training; to answer any questions that arose from the training; to review the BARD gold-standard solution for Problem A and answer any questions; and to conduct a similar review for Problem B. These "gold-standard solutions"" were simply plausible example solutions we constructed, including the associated BNs, that would have achieved the maximum possible rubric score. PowerPoint slides and BARD walkthroughs were used to explain these solutions and how to use BARD to develop them[25]. No webinar was conducted after Problem C, as there were no subsequent problems where participant performance could benefit from further retention or training; however, participants were sent the gold-standard solution via email.

Assignment to teams and roles: X participants were permanently assigned to one of six timeslots spread over three weekday evenings, consistent with their stated availability and our capacity, and asked to keep this timeslot free for participation throughout the experiment. They were then

[25] "Training Presentation - Group X" available at: https://bit.ly/2VnjW1d.

randomly assigned to BARD teams within this timeslot before each problem cycle. Reassignment was another modification to cope with attrition, by maintaining participant numbers within each team. Initially, there were 25 teams made up from 198 people selected for the X condition, but attrition reduced the number of teams across the experiment (see section 5.1). Teams were assigned 6–9 members (except for one team of 5 during the final problem) with an average of 7.3 members for the experiment. We expected some attrition within teams during each problem, i.e., that not all assigned members would actively participate, so the numbers assigned were slightly generous.

As described in section 3, each BARD team had one facilitator and the remainder were analysts. The prospective facilitators were assigned to teams first and distributed as evenly as possible, since BARD includes functionality for facilitators to be replaced. Within each team, the participant with the highest score on the quiz done at registration was selected as the facilitator.

Workflow: BARD's workflow was designed to allow asynchronous problem solving, i.e., with no real-time communication. However, to increase social engagement, team members in this experiment were required to work synchronously online during their allocated 2-h sessions, which was feasible because almost all participants were within the local AEST timezone. While lab-based experimentation would have been even better for combating attrition, as used in Cruz et al. (2020), here our resources were insufficient. Once a problem was "opened" at the start of the team's scheduled session, the participants still had access to BARD and the problem for the remainder of the week until it was "closed" at midnight on Sundays, and so could continue to work on it after the scheduled session time, albeit without additional compensation. In practice, while some participants continued to work on the solution the same night, no participants came back on subsequent days.

Report submission: When the problem was closed, the rules for report submission were:

1. If the facilitator has already submitted a final report, that report will be assessed. The facilitator was trained and instructed to produce the report by either:
    (a) incorporating elements from any or all of the individual reports, or
    (b) choosing what appears to be the best analyst report, based on team consensus via the discussion forum and/or ratings[26].
2. If the facilitator does not submit a report, then among those reports given a rating by at least two analysts, BARD auto-submits the one with the highest mean rating.
3. If there is no report rated by at least two analysts, then BARD auto-submits the longest non-blank analyst report[27].

4.3.3. Control Condition (K)

Training: K individuals received webinars and upfront training for their own tools that were as similar to X as practical[28]. Nevertheless, the content between the X and K webinars differed significantly. K used Google Suite, and their upfront training consisted of an e-course developed by IARPA called the "Guide to Good Reasoning"[29], which provided generic training on how to reason and solve problems, including avoiding the common analytic errors IARPA had already identified.

Webinars followed the same week-by week schedule described in **Table 1** and had similar aims. Each webinar was presented three times within each week, and individuals nominated the timeslot they preferred at the beginning of the experiment. In the webinars following Problems A and B, we presented versions of the gold-standard solutions with almost identical text to those for X, but stripped of any allusion to the BARD tool. We used PowerPoint slides[30] to introduce and explain how "frequency formats" and "chain event graphs" could be used to accurately calculate the answers (see Gigerenzer and Hoffrage, 1995), and also how the elementary probability calculus could be used as a supplement or alternative method, albeit more mathematical and less intuitive. These are the best available pen-and-paper techniques for probability calculation, and were sufficient, in principle, for solving all our problems precisely.

The main motivation for presenting these techniques was to encourage continued participation. As discussed in section 2, the more "ecologically valid" and favorable comparator would have been individual analysts working on problems without any special training in probability calculation, as used in Cruz et al. (2020). For intelligence analysis, these pen-and-paper techniques aren't part of business as usual, and moreover, are not a viable alternative to BNs: although a feasible low-tech alternative for these simplified test problems that require computing a few explicit and precise probabilities, they rapidly become too unwieldy and difficult as problems become more complex or vague. Nevertheless, although we expected this training to improve the performance of K, we reasoned that if X could still outperform K here, then it would outperform an untrained K by at least as great a margin.

Workflow: For ecological validity, K participants worked on each problem individually. A welcome side effect was that it allowed us to maintain the study's statistical power despite limited funding for participant compensation. Analysts in K were provided with individual Google Drive folders containing the Good Reasoning Guide, and for each week the relevant problem statement and blank "Answer Document." K had 58 participants initially, with 51 completing training, and further

[26] Admittedly, there was no active monitoring or intervention in this experiment to prevent facilitators from flouting this training by industriously building their own independent BN solution and writing a report based solely on it—but we received no complaints from analysts that such dictatorial behavior occurred.

[27] For the IARPA experiment, we defined a similar set of rules to classify and submit a report as "non-deficient," and slightly stricter rules requiring participation from several analysts (per the intended social process) to classify it as "Ready-to-Rate." The latter would have been a better basis for assessment if the sample sizes had been sufficient. "BARD Report Flags" available at: https://bit.ly/2CTk3u7.

[28] "Control Group Plan" available at: https://bit.ly/2YOkEXo.

[29] "Guide to Good Reasoning" available at: https://bit.ly/2WJtpQJ.

[30] "Training Presentation - Group K" available at: https://bit.ly/2IduCMo.

attrition leading to smaller numbers for problem-solving (see section 5.1).

Report submission: For K, problems were "opened" on Monday simply by releasing the problem description, and participants had the entire week to work on their report at their convenience. They were free to enter their solution in the Google Drive anytime between the opening and the close on Sunday at midnight, and any non-blank Answer Document was assessed.

4.3.4. Marking

Six markers were engaged with proven marking ability: fluency in English and a background in academic marking. Marker training included a review and discussion of an Assessment Guide[31], as well as a joint session marking example reports. Markers were trained to adhere as closely as possible to a literal interpretation of the problem rubrics and ignore redundant information. Markers were obliged to work independently of each other and BARD project members. Reports were anonymized and marking done blind; in particular, markers were not informed whether they were marking an X or K report.

Markers could not be kept completely blind, however, since only the BARD reports were generated using a structured template, with encouragement to include BARD graphics. As discussed in section 3.4, these are beneficial features of BARD, both because they remind users to provide some oft-neglected content and because they help to present that content more clearly. The potential problem here is not that markers might give legitimate rubric points for providing such content, but rather, that they might become biased in their interpretation of which reports are providing it, and hence illegitimately award points to X or not award points to K. Fortunately, the items awarded rubric points are all very specific pieces of information and it is difficult to misinterpret whether these are provided. However, we endeavored to minimize any such bias by explicitly urging markers to avoid it, and informing them that their performance would be tested for it: some fully anonymized K reports would be camouflaged to appear as X reports and vice versa.

4.4. Statistical Design

The design was pre-registered with The Open Science Framework (OSF)[32], and in accordance with our IARPA contract, stated that inferences about our main hypothesis would be primarily based on 80% and 95% confidence intervals (CIs) for condition means, and standardized effect sizes. We proposed to show that X had the higher mean rubric scores overall (across all three problems), with favorable non-overlapping CIs taken as confirmation of the hypothesis. We also present below the results of some more usual null hypothesis significance tests.

We did not explicitly set a precise target effect size. CREATE, however, had specified at the outset its own performance goals for "Quality of Reasoning" to be achieved by the end of each of its three Phases: Cohen's d (pooled) of at least 0.25 (small), 0.5 (medium), and 1.0 (large) respectively. $d \geq 1.0$ was an ambitious final target, since for structured analytic techniques, this is a major effect that has rarely been robustly achieved. For example, "Argument Mapping" (AM) is a well-known software-supported structured technique where an analyst makes a non-causal, non-parameterized tree diagram to illustrate the logical structure of an argument. A meta-analysis by Alvarez Ortiz (2007) showed that, at best, a one-semester university course using AM improved student critical thinking scores by approximately 0.6 Cohen's d compared to other courses. If $d \geq 1.0$ could be achieved (e.g., in Phase 3), it would undoubtedly be of practical importance. At this stage of BARD's development (Phase 1), IARPA considered $d \geq 0.25$ substantial enough to warrant further funding and development (Phase 2).

There is no natural scale for measuring reasoning performance, so the use of standardized effect size measures that are relative to observed variability is appropriate. But IARPA's blanket specification of Cohen's d as the standardized effect size measure was not optimal, and we pointed out some beneficial refinements. Cohen's d measures effect size in units of observed standard deviation (SD), and calculates this by pooling the SD of K and X. Better is Hedges' g, which also pools the SDs of K and X but corrects for a bias in Cohen's d where group sizes are small and unequal. For CREATE's purposes, better still is Glass' Δ, which uses only the SD of K. That's because, (i) "business as usual" is the relevant norm, and (ii) each new structured method is quite likely to have a different SD, and (iii) "business as usual" is therefore the only common standard of comparison for the four diverse methods. As Glass argued (Hedges and Olkin, 1985), if several treatments are compared to the control group, it's better just to use the control SD, so that effect sizes won't differ under equal means and different variances. Preserving the validity of the comparison in this way outweighs the slightly reduced accuracy of the estimation. An additional consideration is that for ANOVA-based analyses, it is usual to use a proportion of variance explained, e.g., by using partial eta-squared (η^2) rather than Cohen's d. Accordingly, we report our effect sizes via two alternative measures below, but our preferred measure is Glass' Δ.

For the analysis of Problem A, an independent samples t-test was selected to assess the mean difference in rubric scores between K individuals and X teams, along with CIs for the two condition means. Effect size was reported using both Hedges' g and Glass' Δ.

For the analysis of the repeated measures data in Problems B and C, mixed-model ANOVA tests were selected to determine whether any difference in rubric scores is the result of the interaction between the "type of treatment" (i.e., membership of K or X) and "experience" (i.e., solving Part 1 or Part 2) alongside individual main effects for treatment and experience. Where the interaction term was not significant, rubric score differences between K and X were assessed through main effects for the type of treatment, and where the interaction term was significant, through the statistical significance of the simple main effects. Differences were computed using 80 and 95% two-sided Cousineau-Morey confidence intervals for condition means, and the 95% intervals were illustrated graphically. Effect size was reported using both partial η^2 and Glass' Δ.

[31] "Assessment Guide" available at: https://bit.ly/2G1YUjD.

[32] "Experiment Design" available at: https://bit.ly/2OBVBCc.

TABLE 2 | Attendance by week: task completions and both week-on-week and end-to-end attrition, for K individuals, X individuals, all individuals, and X teams (along with the mean attendance per team).

Condition	Attendance	Registration	Training	W1	W2	W3	W4	W5	Weekly	End-to-end
K	Completed	58	51	44	34	31	28	28	–	–
K	Attrition %	–	12%	14%	23%	9%	10%	0%	11%	52%
X	Completed	198	140	130	122	112	114	105	–	–
X	Attrition %	–	29%	7%	6%	8%	−2%	8%	9%	47%
K+X	Completed	256	191	174	156	143	142	133	–	–
K+X	Attrition %	–	25%	9%	10%	8%	1%	6%	10%	48%
X Teams	Completed	25	25	25	23	23	22	21	–	–
X Teams	Attrition %	–	0%	0%	8%	0%	4%	5%	3%	16%
per X Team	Completed	–	5.6	5.2	5.3	4.9	5.2	5.0	5.2	–

To explore potential marker bias due to report formatting, in each of the 5 problem-solving weeks we took three X and three K reports from participants and camouflaged them as reports from the opposing condition. We then randomly presented some blinded markers with the originals and others with the camouflaged versions. To analyze these 30 matched pairs of rubric scores, we used a mixed effects model with fixed effects (for condition and camouflage) and participant level random intercepts to test for any major bias.

5. RESULTS

5.1. Attrition, Missing Values, and Bias

Attendance statistics for individual participants are shown in **Table 2**. To measure end-to-end attrition, the initial numbers are all participants who completed registration and confirmed their availability, and the final numbers are all participants who completed the task in Week 5. End-to-end attrition was about 50% in both conditions, although it was slightly lower (i.e., attendance was slightly better) in X than in K.

Intermediate attendance numbers reveal that week-on-week attrition averaged about 10% in both conditions, although slightly lower in X than K, and tended to reduce as the experiment progressed. A notable difference between conditions is that in K the attrition during training was similar to subsequent problem-solving weeks, whereas in X the attrition during their more substantial training was much higher than K (more than double), but in subsequent weeks was almost always lower than K. Since all trained participants were allowed to resume participation even if they missed a week of problem-solving, it was possible for week-on-week attrition to be negative, which did occur when more X participants completed their task in Week 4 than Week 3.

In terms of teams, the number of X individuals available at each randomized allocation was sufficient to form 25 teams after registration, 23 before Problem B, and 22 before Problem C, with a mean size of 7.3 members for the experiment. Individual attrition resulted in a mean size of 5.2 members actively participating each week, which we expected would be sufficient for the BARD social process to confer significant benefits. Every team completed all of their weekly problem-solving tasks, except for one team in the final week, so 25, 23, and 21 X teams completed Problems A, B, and C, respectively. This equates to an end-to-end attrition for teams of only 16%, and a mean week-on-week attrition rate of only 3%. These are one third of the rates for X individuals, because the rest of the individual attrition occurred within teams[33].

TABLE 3 | X report submission method by week.

Submitted by	W1	W2	W3	W4	W5	All
Facilitator	25	16	15	18	17	91
Automation	0	7	8	4	4	23
Total	25	23	23	22	21	114

TABLE 4 | Individual attendance at optional feedback sessions.

Condition	W1	W2	W4	All
K	32	29	31	92
X	129	115	106	350
Total	161	144	137	442

Since reports from Weeks 2 and 3 were analyzed collectively as part of the phased Problem B, only matched pairs were included: any report from either of these weeks was regarded as an incomplete datum and discarded if the K individual or X team did not also produce a report for the other week. Reports from Weeks 4 and 5 were treated similarly. Fortunately, in K individual attrition was noticeably reduced in the second week of a phased problem. This was not true for individuals in X, but as noted, only one such incomplete datum was produced by X teams. Missing data from X teams was also reduced via our automatic submission contingency: 20% of the X reports assessed were auto-submitted by the BARD system after the facilitator failed to submit, as shown in **Table 3**.

Attendance at the optional, compensated webinars was very good for K and excellent for X, as shown in **Table 4**. Relative to attendance the previous week for the associated upfront training or problem, webinar attendance was 73% in K and 92% in X.

[33] "Attrition Rates" available at: https://bit.ly/2K7wFnG.

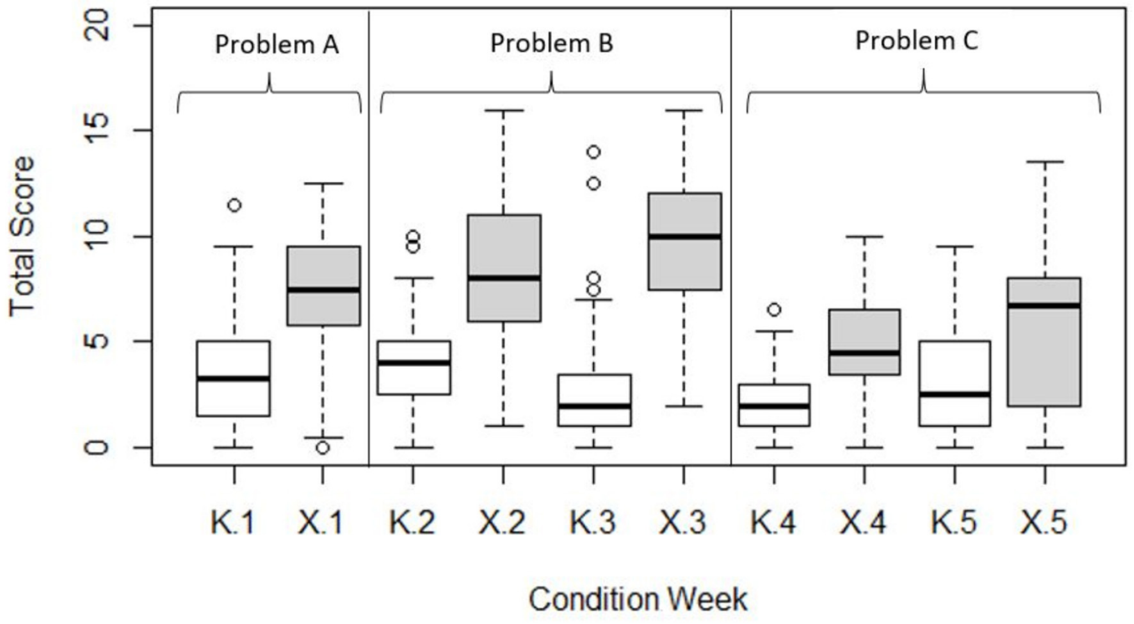


FIGURE 3 | Boxplot of score quartiles for each condition in each problem-solving week. Medians are represented by the thicker horizontal bars. Outliers are represented by circles, and defined as further than 1.5 times the interquartile range from their condition median.

TABLE 5 | Mean scores with their 80% and 95% CIs for each condition in Problem A.

Cond.	Week	N	Mean	SD	SE	95% CI	80% CI	Max
K	1	44	3.545	2.283	0.344	[2.851–4.240]	[3.097–3.993]	13
X	1	25	7.370	2.775	0.555	[6.225–8.516]	[6.639–8.101]	13

Comparing the 30 camouflaged reports to their original counterparts, we did not detect any effect of the report format on the rubric scores awarded by markers [$\chi^2(1) = 0.143$, $p = 0.706$].

5.2. Test Assumptions

For each problem set, assumptions of normality and homogeneity of variances were assessed using Shapiro-Wilk and Levene tests respectively, applied across repeat-condition subgroups and assessed at 95% confidence.

The Shapiro-Wilk test rejected the null of normality only for K in Part 2 of Problem B ($p < 0.001$) and K in Part 1 of Problem C ($p = 0.049$), so these were further assessed using normal quantile-quantile (QQ) plots[34]. The QQ plot for K scores in Part 2 of Problem B was approximately normal, but revealed a single outlier individual performing well above the rest of K, and its temporary removal resulted in an acceptable Shapiro-Wilks outcome ($p = 0.099$). Anecdotal evidence from marked reports suggested that, contrary to the experimental guidelines, some of the highest performers in K used Bayesian analysis methods or tools (other than BARD) to produce their solutions. Such individuals will have increased the mean scores in K. However, we are averse to permanently removing this particular outlier and other unanticipated observations, especially given that these observations favor K. Also, it is well-known that ANOVA can tolerate data that is non-normal, and simulation studies using a variety of non-normal distributions have shown that the false positive rate is not affected substantially by violation of this assumption under an approximately normal distribution (Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996). Visual inspections of QQ plots for K scores in Part 1 of Problem C indicate that, again, the distribution is sufficiently normal to allay concerns about inflated false positive rates.

Levene's test for homogeneity of variances was not significant for week–condition sub-groups in Problem A [$F_{(1,\ 69)} = 1.319$, $p = 0.255$] or Problem B [$F_{(3,\ 100)} = 1.112$, $p = 0.348$], but was significant for Problem C [$F_{(3,\ 84)} = 5.406$, $p = 0.002$]. Kim and Cribbie (2018) show that the impact of departures from homogeneity on false positive error rates are limited when sample sizes are close to equal. The sample sizes in Problem C were 23 for K and 21 for X, so we anticipate this departure from the homogeneity assumption will also have limited impact on false positive rates in final outcomes.

5.3. Analysis

Figure 3 is a set of box plots summarizing and exploring the quartile distribution of rubric scores, after balancing data for attrition by dropping observation pairs with missing values. Median scores indicate that the "middle" team in X always outperformed the "middle" individual in K in all weeks, providing some initial support for our main hypothesis. While the higher

[34] "Quantile Quantile Plots" available at: https://bit.ly/2YMHYot.

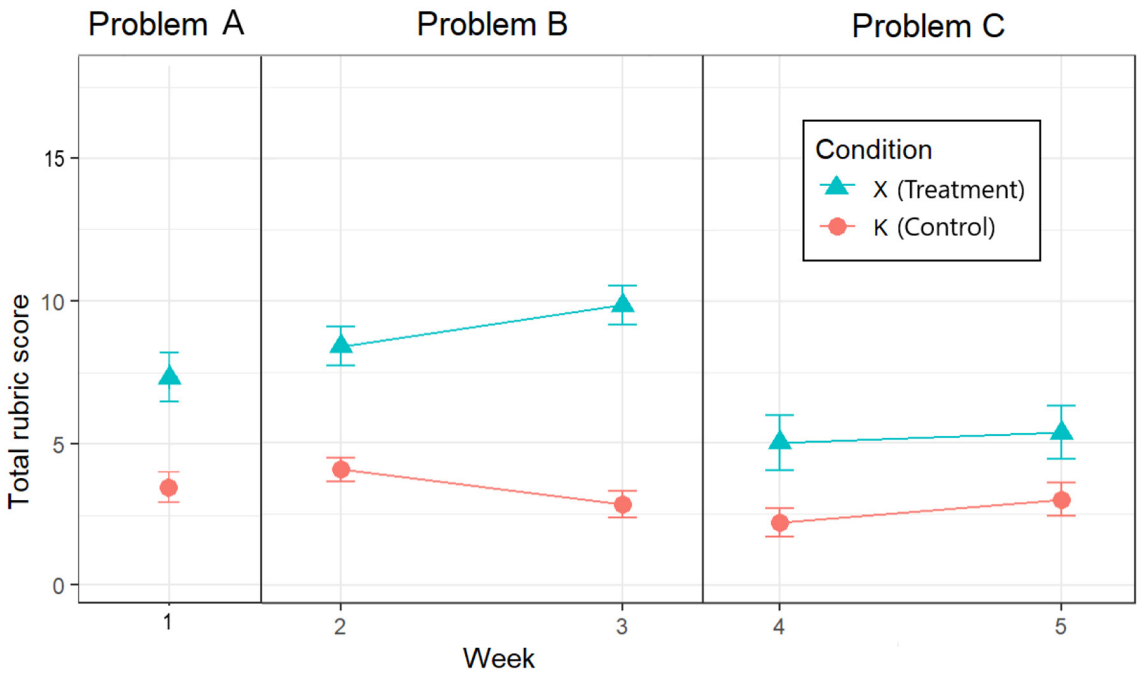


FIGURE 4 | Mean scores with their 95% CIs for each condition in each problem-solving week.

quartiles of K predominantly overlap with the lower quartiles of X, as noted, there are a few outliers in K that perform as well as their high performing X counterparts. Median scores in K vary little across weeks. The superiority of the medians in X is most striking for Problems A and B, and somewhat less for Problem C.

5.3.1. Problem A: Smoking and Cancer

For Problem A, the difference in mean scores between X and K was statistically significant [$t_{(91.223)} = 7.799, p < 0.001$] using an independent samples (Welsch) t-test. 80% and 95% confidence intervals were calculated around each condition's mean score (see **Table 5** and **Figure 4**) and do not overlap, further indicating significantly higher mean scores in favor of BARD.

Given the unequal sample size (K = 44, X = 25), we computed the adjusted Hedges' g effect size of 1.44, while Glass' $\Delta = 1.6$. On either measure, this is considered a very large effect.

5.3.2. Problem B: Kernel Error

For Problem B, inspecting the Week × Condition mean rubric scores for Weeks 2 and 3 depicted in **Figure 4**, we can see that the difference between the control and experimental conditions increases, which suggests a Week × Condition interaction. Indeed, our 2 × 2 mixed ANOVA showed a statistically significant interaction between experimental condition and problem week [$F_{(1,\ 50)} = 8.93, p < 0.001$]. The main effect of experimental condition was significant [$F_{(1,\ 50)} = 86.46, p < 0.05$], while the mean effect of exposure week was not [$F_{(1,\ 50)} = 0.06, p = 0.81$].

Adjusted confidence intervals as described by Morey (2008) were calculated around each Week × Condition mean score (see **Table 6** and **Figure 4**), and do not overlap for K and X in either Week 2 or Week 3, further indicating significantly higher mean scores in favor of BARD.

The size of the main effect of condition as measured by the generalized η^2 is 0.53, which is considered very large (Bakeman, 2005), while Glass' $\Delta = 2.2$, which is considered huge. The generalized η^2 effect size for the Week × Condition interaction, 0.06, is considered small[35].

5.3.3. Problem C: Cyberattack

For Problem C, inspecting the Week × Condition mean rubric scores for Weeks 4 and 5 depicted in **Figure 4**, we can see that the difference between the control and experimental conditions is similar, which suggests no Week × Condition interaction. Indeed, our 2 × 2 mixed ANOVA showed there was no statistically significant interaction between experimental condition and problem week [$F_{(1,\ 42)} = 0.35, p < 0.56$]. The main effect of experimental condition was significant [$F_{(1,\ 42)} = 17.68, p < 0.05$], while the main effect of exposure week was not [$F_{(1,\ 42)} = 2.58, p = 0.12$].

Again, adjusted confidence intervals were calculated around each Week × Condition mean score (see **Table 7** and **Figure 4**), and do not overlap for K and X in either Week 4 or Week 5, further indicating significantly higher mean scores in favor of BARD.

The size of the main effect of condition as measured by the generalized η^2 is 0.24, which is considered large (Bakeman, 2005), while Glass' $\Delta = 1.4$, which is considered very large.

[35] "ANOVA Tables and Main Effects" available at: https://bit.ly/2YPRDec.

TABLE 6 | Mean scores with their 80% and 95% CIs for each condition in Problem B.

Cond.	Week	N	Mean	SD	SE	95% CI	80% CI	Max
K	2	29	4.069	2.437	0.320	[3.428–4.710]	[3.654–4.484]	18
X	2	23	8.402	3.575	0.527	[7.341–9.464]	[7.717–9.088]	18
K	3	29	2.845	2.721	0.357	[2.129–3.560]	[2.382–3.308]	16
X	3	23	9.848	3.591	0.529	[8.781–10.914]	[9.159–10.536]	16

TABLE 7 | Mean scores with their 80% and 95% CIs for each condition in Problem C.

Cond.	Week	N	Mean	SD	SE	95% CI	80% CI	Max
K	4	23	2.208	1.751	0.253	[1.700–2.717]	[1.880–2.537]	22
X	4	21	5.012	3.070	0.474	[4.055–5.968]	[4.395–5.629]	22
K	5	23	3.022	1.972	0.291	[2.436–3.607]	[2.643–3.400]	16
X	5	21	5.369	3.009	0.464	[4.431–6.307]	[4.764–5.974]	16

6. DISCUSSION

6.1. Likely Causes and Effects of Attrition

For Delphi studies, which necessarily require participants to respond for two or more rounds on the same test problem, attrition rates per round can be high, accumulate to extremely high levels, and threaten to bias the results (Toma and Picioreanu, 2016). Some typical *initial* attrition rates (i.e., at the second round) reported in the literature are approximately 15% (Elwyn et al., 2006), 30% (Bradley and Stewart, 2002), and 50% (Moreno-Casbas et al., 2001; Goluchowicz and Blind, 2011). In comparison, the 10% weekly attrition rates we achieved were very low, and our end-to-end rate of 50% after six rounds was, although high, about equal to the attrition rate seen after one round in the latter studies and our piloting. Consequently, unlike the larger IARPA study, we managed to cater for and reduce individual attrition sufficiently to obtain statistically significant results, assisted by our participant compensation, social engagement, team sizes, and auto-submission.

In the aborted IARPA study, in which four contrasting systems were being tested along with a control condition similar to our own, the length of upfront training varied between systems, and attrition rates during this training were roughly proportional to its duration. Since upfront training was at least twice as long in X as in K, the doubled attrition in X compared to K is consistent with the IARPA study, and does not imply that our training was particularly difficult.

In the problem-solving weeks, two possible causes for the slightly lower individual attrition in X compared to K are benefits of working as a group: the mean task burden per individual can be reduced by distributing it (often not evenly!) amongst team members, and the social interactions involved in the task can make it more attractive. Another possible cause is a benefit of using BNs: the tool may have seemed better suited to the task, encouraging participants to persist with it.

We expect the main introduced bias due to attrition was that participants who felt more competent at the task were more likely to show up for subsequent rounds, potentially improving average performance in the condition. Performance in both K and X may have progressively benefited from this, but the principal concern here is that they may not have benefited equally, thus contributing to our effect sizes one way or the other. Both K and X involved using techniques (mathematical and modeling respectively) that some individuals would have been able to use better than others, so in this respect it isn't clear which condition's individuals would have benefited more. There are, however, two social factors that clearly should have reduced the benefit to X teams. First, participants who felt less competent should already have had less impact than their team members on the team report, so their absence probably didn't improve the team responses as much per report as attrition in K. Second, in a group social process like Delphi, less capable participants may still make a positive net contribution to a group report, so it is possible that their absence actually made X reports worse. Finally, although it may have been obscured by the variation in problem-solving tasks, there was no observable trend of increasing effect sizes over the 5 weeks as the level of attrition increased. For all these reasons, it seems very unlikely that attrition made a major positive contribution to our headline result: that X consistently outperformed K.

There is one other, important reason why IARPA, at least, was sanguine about possible attrition bias. The intended use of any CREATE system was not to make it a compulsory tool, replacing business as usual for all analysts. Rather, it was to make it available as an optional alternative for any analysts who are attracted to it and voluntarily persist with it. This was, similarly, expected to select those who feel more competent using the system, and create a self-selection and attrition bias far greater than any in our experiment. Hence, although we strived to minimize attrition, any attrition bias there may have been in our experiment will only have made our results a more accurate indicator of likely performance for IARPA's intended use.

6.2. Effects of Problem Difficulty

We know that much more complex problems of a similar kind can be solved accurately by BN experts using the tools in X, and aren't tractable for anyone using the tools in K. So, we expected that the increase in complexity in the second phase of Problems B and C would translate into a bigger advantage for X over K. However, the advantage detected in B was small, and not detected at all in C. It may be that the increases in complexity and/or the ability of our participants to use BN models to overcome it was not as great as we supposed.

Since the second phase of Problem B involves the "explaining away" cognitive difficulty, whereas the second phase of Problem C involves the "dependent evidence" cognitive difficulty, this may suggest that explaining away is more difficult to understand than dependent evidence. However, this interpretation would be unwarranted. We used only one example of each difficulty, so there are numerous confounds; and this effect-size ordering was not observed in the SoloBARD experiment. Measuring the relative difficulty of various cognitive difficulties would require many further, more careful comparisons.

6.3. Robustness and Size of Effect

The superior performance of X over K was a robust effect across our three problems, since it was confirmed independently for each. On our preferred measure, the effect sizes were all very large to huge (Glass' Δ 1.4–2.2), and their 95% CIs are shown graphically by week in **Figure 5**. On any standard measure, they greatly exceeded CREATE's initial target of a small effect size, and indeed, achieved in Phase 1 for simple problems the large effect size desired in Phase 3 for more complex Problems.

It is interesting to compare the performance of our participants to those in the SoloBARD experiment, which used a similar set of three problems (a different problem instead of our Problem A, but exactly the same Problems B and C). Unexpectedly, SoloBARD participants performed better than ours. However, this was mainly in the control condition—so, as expected, our BARD users beat our controls by a greater margin than the SoloBARD users beat their controls. Specifically, SoloBARD control individuals performed much better (obtaining 32% of the available points) than our control individuals (obtaining only 17%), while SoloBARD experimental individuals performed only slightly better (obtaining 48%) than our BARD experimental teams (obtaining 41%). Consequently, BARD achieved double the mean effect size (Glass' $\Delta = 1.7$) of SoloBARD (Glass' $\Delta = 0.8$).

There were multiple differences between the two experiments that may have affected performances, so we must be cautious in attributing specific causes to the differences in results. However, we see no factor likely to have benefited *only* the SoloBARD controls compared to our K. On the contrary, while in both experiments control individuals received IARPA's Guide to Good Reasoning, our K individuals also received some training in pen-and-paper probability calculation techniques, which should have improved their relative performance—yet this effect is not evident, perhaps because it is swamped by other factors. In contrast, there are several plausible causes for better performance in *both* the control and experimental conditions of SoloBARD compared to our K and X: (i) superior ability of participants, who were drawn solely from the University College London experimental participant pool rather than recruited on the more *ad-hoc* basis described in section 4, (ii) in-lab testing rather than online, which tends to improve motivation and compliance, and (iii) offering substantial and extensive financial bonuses for good performance (to supplement a modest hourly rate), rather than just offering a generous hourly rate. It is possible that these factors made more difference to the relative performance of the control conditions than to the experimental conditions. However, there is a more obvious explanation for the greater outperformance of the experimental over the control condition in our BARD experiment: our X participants benefited from working in small groups. This is consistent with the general prior literature on Delphi and our specific prior experiment with Delphi in BARD, as summarized in sections 2.4, 2.5, and 2.7.

In summary, there were clearly significant factors driving down performance in our experiment compared to the SoloBARD experiment, and there may have been an interaction effect that contributed one way or the other to our effect sizes. Nevertheless, with that caveat, the doubled effect size achieved by BARD in comparison to SoloBARD suggests, and provides some cumulative evidence, that our social processes make a substantial positive contribution in addition to the substantial positive contribution made by BN construction.

6.4. Quality of Reports, Causal Models, and Training

As expected, the mean proportions of the available rubric points obtained by participants were low, even when assisted by BARD. As discussed in section 4.2, participants are unlikely to provide a high proportion of the specific items the rubric rewards. Our rudimentary AI tool was designed to suggest possible text to include in justifications, but at this early stage of development it was not able to suggest all the relevant points, and apparently it had limited effect.

Given that our X participants were supposed to achieve better written reports than K by constructing BN models, it is natural to ask how accurate their models turned out to be, and how well-correlated this was to the quality of their written reports. In Bolger et al. (2020), our BARD team members assessed the quality of BN structures by measuring the difference between these and normatively correct "gold-standard" structures using "edit distance," which is the most well-known structural measure in the literature (e.g., Spirtes et al., 2000). However, this approach was facilitated by requiring participants to choose variables out of a set provided, and not requiring probabilities to be entered in the models, thus avoiding both sources of variation in participant answers. Furthermore, the aim was to compare the relative quality of structures produced by individuals before and after peer feedback, not their absolute quality. Here, even if we used a broader measure of the overall quality of the BNs produced by our X teams, there would be no meaningful comparison to evaluate how well our X teams performed. The only way to compare X to K performance is to measure the quality of their written reports, which they both produce, and are designed to implicitly test the accuracy of the BNs constructed by X teams via the accuracy of their answers. The same inherent limitation applied to the SoloBARD experiment. However, we will make all the BNs produced by our X teams available for subsequent research.

For similar reasons, it is difficult to assess the skill level in BN construction achieved by our X teams. There is, as yet, no standard test for BN-modeling difficulty or ability, so we can't quantify more precisely the difficulty of building our problem BNs or the ability achieved through our minimal training. However, it is notable that, as in our two previous experiments, our BARD users received very little BN training by industry standards, and yet they were able to construct BNs well enough to outperform control participants on our probabilistic reasoning problems. This provides some welcome evidence that intelligence analysts, for example, can be quickly trained to use BNs using our online resources. We are also confident that with further training and experience, X teams would substantially improve their BN building skills and consequently their written reports.

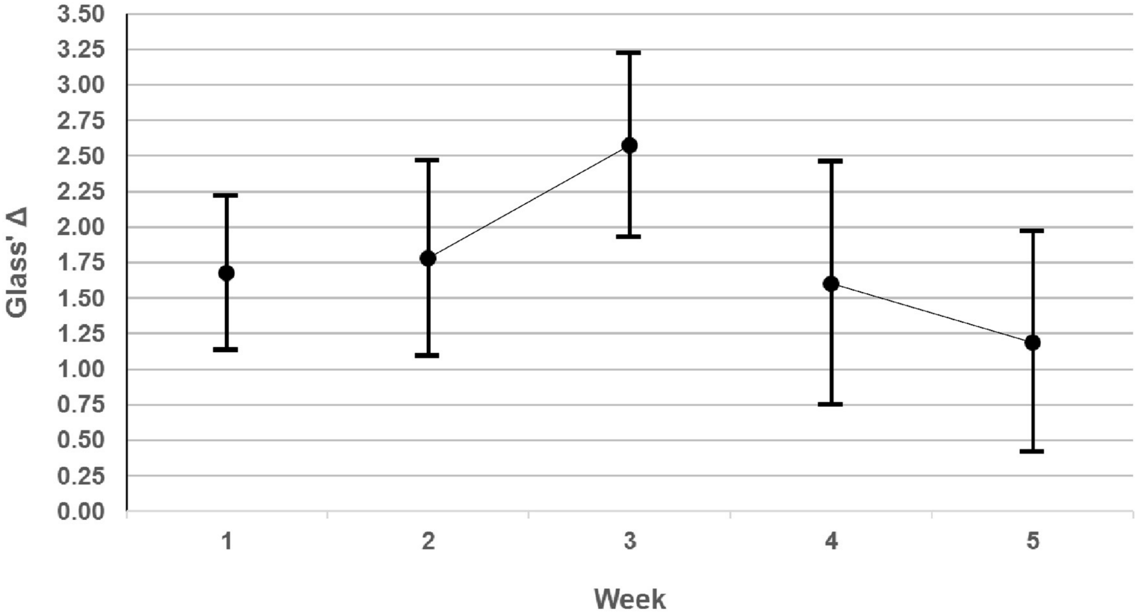


FIGURE 5 | Effect sizes measured by Glass' Δ with their 95% CIs for each week.

7. CONCLUSION

Our results show that BARD is an extremely promising tool for intelligence analysis that warrants further research. Compared to business as usual, it already performs much better on simple test problems. Compared to existing BN software, it offers a unique integration of BN construction with a Delphi-style collaborative workflow, high-quality online training and help, and a structured template for written reports with complementary text explanations automatically generated from the BN. Furthermore, there is enormous potential for further research and improvement: in developing more complex problems, in developing BARD's features, and in testing their individual and combined efficacy on those problems. There are also numerous potential applications for BARD outside intelligence analysis, since many areas—including those to which BNs have already been introduced—require reasoning and decision making under uncertainty.

More generally, our results provide some cumulative evidence (in addition to prior theory and experiments) for the utility of BARD's key components:

- Good online training allows people who are not BN experts to construct BNs, minimizing the need for a facilitator who is a BN expert.
- Where time permits, BN construction can be used effectively for probabilistic reasoning problems. This helps to avoid numerous types of causal and probabilistic reasoning difficulties, and adds precision.
- Small group collaboration, via RT Delphi in particular, can be used successfully for BN construction. This allows multiple viewpoints to be debated and combined to produce a better result.

Three issues for further research deserve particular emphasis:

1. We must test the efficacy of probability estimation. Our team showed that it is possible and necessary to develop a new type of test problem for probabilistic reasoning: sufficiently challenging, yet simple enough to assess (with many normatively correct elements in the solution). More complex problems of this sort must be developed that include the estimation of probabilities by experiment participants, rather than relying entirely on precise parameters specified in the problem statement. BARD's built-in capacities for eliciting and combining probability estimations can then be rigorously tested.
2. Our social processes, in addition to RT Delphi, include components such as discussion boards and the rating of other team members' work. These components can be evaluated and optimized individually and in combination. If such components work sufficiently well, then in many applications BARD could dispense with the human facilitator altogether without much loss.
3. Our automated verbal explanations were novel and promising, but we have not yet measured their contribution. Moreover, we now believe this XAI tool would be better implemented as a combination of visual and verbal features that are more interactive. Our spin-off project, mentioned in section 3.4, will investigate this in detail[36].

[36] "Improving human reasoning with causal Bayes networks: a multimodal approach." See https://dataportal.arc.gov.au/NCGP/Web/Grant/Grant/DP200100040.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on OSF or on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Human Research Ethics Committee, Monash University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All of the authors were involved in the design, construction and testing of the BARD application to greater or lesser extents, excepting MO and YL. The latter two as well as AO were involved in data analysis, with MO (from Statistics, Monash) doing much of the heavy lifting. RP was involved in the organization and day-to-day oversight of the experiments, while ST and AO organized participants, ran the BARD webinars and were online during the synchronous BARD sessions for technical support as required. KK and EN conducted the control group webinars. KK led the design and oversaw the running of the experiment, as well as leading the BARD project as a whole and wrote much of this paper. AO drafted the manuscript. AN made significant rewrites. In response to referee feedback, EN managed supplementary data analysis and reporting, and made extensive revisions and additions to the manuscript. All authors reviewed drafts, providing feedback, and suggesting edits.

FUNDING

Funding for the BARD project, and specifically for running the experiment and analyses reported in this paper, was provided by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through their CREATE program under Contract [2017-16122000003][37].

ACKNOWLEDGMENTS

The authors thank Belinda Lewis (BARD Project Officer) for assisting with participant recruitment; Abhi Neog, Maryam Anjiri, Michelle Yu, Abida Shahzad, and Somaiyeh MahmoudZadeh for assisting us by marking the reports; and the three reviewers for their numerous helpful comments on the manuscript.

[37] The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

Office of the Director of National Intelligence (ODNI) (2015). *Intelligence Community Directive 203 (ICD-203): Analytic Standards*. Washington, DC: United States Government. Available online at: https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf

Alvarez Ortiz, C. M. (2007). *Does philosophy improve critical thinking skills?* Master's thesis. University of Melbourne, Melbourne, VIC, Australia.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods* 37, 379–384. doi: 10.3758/BF03192707

Bayraktar, M. E., and Hastak, M. (2009). Bayesian belief network model for decision making in highway maintenance: case studies. *J. Constr. Eng. Manage.* 135, 1357–1369. doi: 10.1061/(ASCE)CO.1943-7862.0000111

Bolger, F., Nyberg, E. P., Belton, I., Crawford, M. M., Hamlin, I., Nicholson, A., et al. (2020). Improving the production and evaluation of structural models using a Delphi process. *OSF Preprints*. doi: 10.31219/osf.io/v6qsp. [Epub ahead of print].

Boneh, T. (2010). *Ontology and Bayesian decision networks for supporting the meteorological forecasting process* Ph.D. thesis. Monash University, Melbourne, VIC, Australia.

Bradley, L., and Stewart, K. (2002). A Delphi study of the drivers and inhibitors of internet banking. *Int. J. Bank Market.* 20, 250–260. doi: 10.1108/02652320210446715

Charness, G., and Sutter, M. (2012). Groups make better self-interested decisions. *J. Econ. Perspect.* 26, 157–76. doi: 10.1257/jep.26.3.157

Chee, Y. E., Wilkinson, L., Nicholson, A. E., Quintana-Ascencio, P. F., Fauth, J. E., Hall, D., et al. (2016). Modelling spatial and temporal changes with GIS and spatial and dynamic Bayesian networks. *Environ. Model. Softw.* 82, 108–120. doi: 10.1016/j.envsoft.2016.04.012

Choi, K.-H., Joo, S., Cho, S. I., and Park, J.-H. (2007). Locating intersections for autonomous vehicles: a Bayesian network approach. *ETRI J.* 29, 249–251. doi: 10.4218/etrij.07.0206.0178

Cruz, N., Desai, S. C., Dewitt, S., Hahn, U., Lagnado, D., Liefgreen, A., et al. (2020). Widening access to Bayesian problem solving. *Front. Psychol.* 11: 660. doi: 10.3389/fpsyg.2020.00660

Elwyn, G., O'Connor, A., Stacey, D., Volk, R., Edwards, A., Coulter, A., et al. (2006). Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. *BMJ* 333:417. doi: 10.1136/bmj.38926.629329.AE

Etminani, K., Naghibzadeh, M., and Peña, J. M. (2013). DemocraticOP: a Democratic way of aggregating Bayesian network parameters. *Int. J. Approx. Reason.* 54, 602–614. doi: 10.1016/j.ijar.2012.12.002

Fenton, N., and Neil, M. (2000). The "Jury Fallacy" and the use of Bayesian networks to present probabilistic legal arguments. *Math Today* 37, 61–102.

Fenton, N., Neil, M., and Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cogn. Sci.* 37, 61–102. doi: 10.1111/cogs.12004

Flores, M. J., Nicholson, A. E., Brunskill, A., Korb, K. B., and Mascaro, S. (2011). Incorporating expert knowledge when learning Bayesian network structure: a medical case study. *Artif. Intell. Med.* 53, 181–204. doi: 10.1016/j.artmed.2011.08.004

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102:684. doi: 10.1037/0033-295X.102.4.684

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev. Educ. Res.* 42, 237–288. doi: 10.3102/00346543042003237

Goluchowicz, K., and Blind, K. (2011). Identification of future fields of standardisation: an explorative application of the Delphi methodology. *Technol. Forecast. Soc. Change* 78, 1526–1541. doi: 10.1016/j.techfore.2011.04.014

Gopnik, A., Sobel, D. M., Schulz, L. E., and Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Dev. Psychol.* 37:620. doi: 10.1037/0012-1649.37.5.620

Gordon, T., and Pease, A. (2006). RT Delphi: an efficient,"round-less" almost real time Delphi method. *Technol. Forecast. Soc. Change* 73, 321–333. doi: 10.1016/j.techfore.2005.09.005

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

Hahn, U., and Harris, A. J. (2014). What does it mean to be biased: motivated reasoning and rationality. *Psychol. Learn. Motivat.* 61, 41–102. doi: 10.1016/B978-0-12-800283-4.00002-2

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

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., and Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. *J. Educ. Stat.* 17, 315–339. doi: 10.3102/10769986017004315

Hedges, L. V., and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Hemming, V., Burgman, M., Hanea, A., McBride, M., and Wintle, B. (2018a). A practical guide to structured expert elicitation using the idea protocol. *Methods Ecol. Evol.* 9, 169–180. doi: 10.1111/2041-210X.12857

Hemming, V., Walshe, T., Hanea, A., Fidler, F., and Burgman, M. (2018b). Eliciting improved quantitative judgements using the idea protocol: a case study in natural resource management. *PLoS ONE* 13:e0198468. doi: 10.1371/journal.pone.0198468

Heuer, R. J. (1999). *Psychology of Intelligence Analysis*. Washington, DC: Centre for the Study of Intelligence, Central Intelligence Agency. Available online at: https://www.cia.gov/library

Holt, D. V., and Osman, M. (2017). Approaches to cognitive modeling in dynamic systems control. *Front. Psychol.* 8:2032. doi: 10.3389/fpsyg.2017.02032

Intelligence Advanced Research Projects Activity (2016). *Broad Agency Announcement (IARPA-BAA-15-11): Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE)*. Washington, DC: United States Government. Available online at: https://beta.sam.gov/api/prod/opps/v3/opportunities/resources/files/8cc3355752ec4965851bcff7770bb241/download?api_key=null&status=archived&token=

Jarvstad, A., and Hahn, U. (2011). Source reliability and the conjunction fallacy. *Cogn. Sci.* 35, 682–711. doi: 10.1111/j.1551-6709.2011.01170.x

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.

Kim, Y. J., and Cribbie, R. A. (2018). ANOVA and the variance homogeneity assumption: exploring a better gatekeeper. *Br. J. Math. Stat. Psychol.* 71, 1–12. doi: 10.1111/bmsp.12103

Korb, K. (2004). Bayesian informal logic and fallacy. *Informal Logic* 24, 41–70. doi: 10.22329/il.v24i1.2132

Korb, K., and Nicholson, A. (2011). *Bayesian Artificial Intelligence, 2nd Edn*. Boca Raton, FL: Chapman & Hall/CRC Computer Science & Data Analysis; CRC Press.

Korb, K. B., and Nyberg, E. P. (2016). Analysing arguments using causal Bayesian networks. *Bayesian Watch*. Available online at: https://bayesianwatch.wordpress.com/2016/03/30/aaucbn/

Kugler, T., Kausel, E. E., and Kocher, M. G. (2012). Are groups more rational than individuals? A review of interactive decision making in groups. *Wiley Interdiscipl. Rev. Cogn. Sci.* 3, 471–482. doi: 10.1002/wcs.1184

Kushnir, T., Gopnik, A., Lucas, C., and Schulz, L. (2010). Inferring hidden causal structure. *Cogn. Sci.* 34, 148–160. doi: 10.1111/j.1551-6709.2009.01072.x

Lagnado, D. A., Fenton, N., and Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument Comput.* 4, 46–63. doi: 10.1080/19462166.2012.682656

Lagnado, D. A., and Gerstenberg, T. (2017). "Causation in legal and moral reasoning," in *Oxford Handbook of Causal Reasoning*, ed M. R. Waldmann (Oxford: Oxford University Press), 565–602. doi: 10.1093/oxfordhb/9780199399550.013.30

Lagnado, D. A., and Sloman, S. (2004). The advantage of timely intervention. *J. Exp. Psychol. Learn. Mem. Cogn.* 30:856. doi: 10.1037/0278-7393.30.4.856

Laskey, K. B., and Mahoney, S. M. (1997). "Network fragments: representing knowledge for constructing probabilistic models," in *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 334–341.

Laskey, K. B., and Mahoney, S. M. (2000). Network engineering for agile belief network models. *IEEE Trans. Knowl. Data Eng.* 12, 487–498. doi: 10.1109/69.868902

Liefgreen, A., Tešic, M., and Lagnado, D. (2018). "Explaining away: significance of priors, diagnostic reasoning, and structural complexity," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, eds T. Roger, M. Rau, X. Zhu, and W. Kalish (Austin, TX: Cognitive Science Society), 2044–2049.

Linstone, H., and Turoff, M. (1975). *The Delphi Method: Techniques and Applications*. London: Addison-Wesley.

Lix, L. M., Keselman, J. C., and Keselman, H. (1996). Consequences of assumption violations revisited: a quantitative review of alternatives to the one-way analysis of variance f test. *Rev. Educ. Res.* 66, 579–619. doi: 10.3102/00346543066004579

Malcolm, D. G., Roseboom, C. E., Clark, C. E., and Fazar, W. (1959). Application of a technique for research and development program evaluation. *Operat. Res.* 7, 646–649. doi: 10.1287/opre.7.5.646

Mascaro, S., Nicholso, A. E., and Korb, K. B. (2014). Anomaly detection in vessel tracks using Bayesian networks. *Int. J. Approx. Reason.* 55(1 Pt 1), 84–98. doi: 10.1016/j.ijar.2013.03.012

Matsumori, K., Koike, Y., and Matsumoto, K. (2018). A biased Bayesian inference for decision-making and cognitive control. *Front. Neurosci.* 12:734. doi: 10.3389/fnins.2018.00734

Misirli, A. T., and Bener, A. B. (2014). Bayesian networks for evidence-based decision-making in software engineering. *IEEE Trans. Softw. Eng.* 40, 533–554. doi: 10.1109/TSE.2014.2321179

Moore, D. A., and Healy, P. J. (2008). The trouble with overconfidence. *Psychol. Rev.* 115, 502–517. doi: 10.1037/0033-295X.115.2.502

Moreno-Casbas, T., Martín-Arribas, C., Orts-Cortés, I., Comet-Cortés, P., and Investén-isciii Co-ordination and Development of Nursing Research Centre (2001). Identification of priorities for nursing research in Spain: a Delphi study. *J. Adv. Nurs.* 35, 857–863. doi: 10.1046/j.1365-2648.2001.01923.x

Morey, R. D. (2008). Confidence intervals from normalized data: a correction to Cousineau (2005). *Tutor. Quant. Methods Psychol.* 4, 61–64. doi: 10.20982/tqmp.04.2.p061

Mumford, M. D., Blair, C., Dailey, L., Leritz, L. E., and Osburn, H. K. (2006). Errors in creative thought? cognitive biases in a complex processing activity. *J. Creat. Behav.* 40, 75–109. doi: 10.1002/j.2162-6057.2006.tb01267.x

Newell, B. R., Lagnado, D. A., and Shanks, D. R. (2015). *Straight Choices: The Psychology of Decision Making, 2nd Edn*. Hove: Psychology Press.

Nicholson, A., Korb, K., Nyberg, E., Wybrow, M., Zukerman, I., Mascaro, S., et al. (2020). BARD: a structured technique for group elicitation of Bayesian networks to support analytic reasoning. *arXiv* 2003.01207.

Nicholson, A., Woodberry, O., Mascaro, S., Korb, K., Moorrees, A., and Lucas, A. (2011). "ABC-BN: a tool for building, maintaining and using Bayesian networks in an environmental management application," in *Proceedings of the 8th Bayesian Modelling Applications Workshop, Vol. 818* (Barcelona), 331–335. Available online at: http://ceur-ws.org/Vol-818/

Nicholson, A. E., Mascaro, S., Thakur, S., Korb, K. B., and Ashman, R. (2016). *Delphi Elicitation for Strategic Risk Assessment*. Technical Report TR-2016, Bayesian Intelligence Pty Ltd. Available online at: https://bayesian-intelligence.com/publications/TR2016_1_Delphi_Elicitation.pdf

Packer, D. J. (2009). Avoiding groupthink: whereas weakly identified members remain silent, strongly identified members dissent about collective problems. *Psychol. Sci.* 20, 546–548. doi: 10.1111/j.1467-9280.2009.02333.x

Pearl, J. (1998). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J., and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect, 1st Edn*. New York, NY: Basic Books, Inc.

Pilditch, T., Hahn, U., and Lagnado, D. (2018). "Integrating dependent evidence: naïve reasoning in the face of complexity," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, eds T. Roger, M. Rau, X. Zhu, and W. Kalish (Austin, TX: Cognitive Science Society), 884–889.

Pilditch, T. D., Fenton, N., and Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychol. Sci.* 30, 250–260. doi: 10.1177/0956797618818484

Pollino, C., Woodberry, O., Nicholson, A., Korb, K., and Hart, B. T. (2007). Parameterisation of a Bayesian network for use in an ecological risk management case study. *Environ. Model. Softw.* 22, 1140–1152. doi: 10.1016/j.envsoft.2006.03.006

Ropero, R. F., Nicholson, A. E., Aguilera, P. A., and Rumí, R. (2018). Learning and inference methodologies for hybrid dynamic bayesian networks: a case study for a water reservoir system in andalusia, spain. *Stochast. Environ. Res. Risk Assess.* 32, 3117–3135. doi: 10.1007/s00477-018-1566-5

Rowe, G., Wright, G., and Bolger, F. (1991). Delphi: a reevaluation of research and theory. *Technol. Forecast. Soc. Change* 39, 235–251. doi: 10.1016/0040-1625(91)90039-I

Russ, S., Rout, S., Sevdalis, N., Moorthy, K., Darzi, A., and Vincent, C. (2013). Do safety checklists improve teamwork and communication in the operating room? A systematic review. *Ann. Surg.* 258, 856–871. doi: 10.1097/SLA.0000000000000206

Salerno, J. M., Bottoms, B. L., and Peter-Hagene, L. C. (2017). Individual versus group decision making: Jurors' reliance on central and peripheral information to evaluate expert testimony. *PLoS ONE* 12:e0183580. doi: 10.1371/journal.pone.0183580

Serwylo, P. (2015). *Intelligently generating possible scenarios for emergency management during mass gatherings* Ph.D. thesis. Monash University, Melbourne, VIC, Australia.

Sesen, M. B., Nicholson, A. E., Banares-Alcantara, R., Kadir, T., and Brady, M. (2013). Bayesian networks for clinical decision support in lung cancer care. *PLoS ONE* 8:e82349. doi: 10.1371/journal.pone.0082349

Silberman, L. H., and Robb, C. S. (2005). *Unclassified Version of the Report of the Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction*. Washington, DC: United States Government. Available online at: https://www.govinfo.gov/app/details/GPO-WMD

Soll, J., and Klayman, J. (2004). Overconfidence in interval estimates. *J. Exp. Psychol. Learn. Mem. Cogn.* 30, 299–314. doi: 10.1037/0278-7393.30.2.299

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., and Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Anal.* 30, 512–523. doi: 10.1111/j.1539-6924.2009.01337.x

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search, 2nd Edn*. Cambridge, MA: MIT Press.

Stacey, K., Sonenberg, E., Nicholson, A., Boneh, T., and Steinle, V. (2003). "A teaching model exploiting cognitive conflict driven by a Bayesian network," in *User Modeling 2003*, eds P. Brusilovsky, A. Corbett, and F. de Rosis (Berlin; Heidelberg: Springer), 352–362.

Stettinger, M., Felfernig, A., Leitner, G., and Reiterer, S. (2015). "Counteracting anchoring effects in group decision making," in *International Conference on User Modeling, Adaptation, and Personalization* (Cham: Springer), 118–130.

Straus, S. G., Parker, A. M., and Bruce, J. B. (2011). The group matters: a review of processes and outcomes in intelligence analysis. *Group Dyn. Theor. Res. Pract.* 15:128. doi: 10.1037/a0022734

Stromer-Galley, J., Rossini, P., Kenski, K., Folkestad, J., McKernan, B., Martey, R., et al. (2018). User-centered design and experimentation to develop effective software for evidence-based reasoning in the intelligence community: the trackable reasoning and analysis for crowdsourcing and evaluation (TRACE) project. *Comput. Sci. Eng.* 20, 35–42. doi: 10.1109/mcse.2018.2873859

Toma, C., and Picioreanu, I. (2016). The Delphi technique: methodological considerations and the need for reporting guidelines in medical journals. *Int. J. Public Health Res.* 4, 47–59.

United States Select Senate Committee on Intelligence (2004). *Report on the U.S. Intelligence Community's Prewar Intelligence Assessments on Iraq*. Washington, DC: United States Government. Available online at: https://www.intelligence.senate.gov/sites/default/files/publications/108301.pdf

van der Gaag, L. C., Renooij, S., Schijf, H. J., Elbers, A. R., and Loeffen, W. L. (2012). "Experiences with eliciting probabilities from multiple experts," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (Cham: Springer), 151–160.

Villejoubert, G., and Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. *Mem. Cogn.* 30, 171–178. doi: 10.3758/BF03195278

Welsh, M. B., and Navarro, D. J. (2012). Seeing is believing: priors, trust, and base rate neglect. *Org. Behav. Hum. Decis. Process.* 119, 1–14. doi: 10.1016/j.obhdp.2012.04.001

Conflict of Interest: 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 October 2020
doi: 10.3389/fpsyg.2020.503233

Propensities and Second Order Uncertainty: A Modified Taxi Cab Problem

Stephen H. Dewitt[1]*, Norman E. Fenton[2], Alice Liefgreen[1] and David A. Lagnado[1]

[1] Department of Experimental Psychology, University College London, London, United Kingdom, [2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

The study of people's ability to engage in causal probabilistic reasoning has typically used fixed-point estimates for key figures. For example, in the classic taxi-cab problem, where a witness provides evidence on which of two cab companies (the more common 'green'/less common 'blue') were responsible for a hit and run incident, solvers are told the witness's ability to judge cab color is 80%. In reality, there is likely to be some uncertainty around this estimate (perhaps we tested the witness and they were correct 4/5 times), known as second-order uncertainty, producing a distribution rather than a fixed probability. While generally more closely matching real world reasoning, a further important ramification of this is that our best estimate of the witness' accuracy can and should change when the witness makes the claim that the cab was blue. We present a Bayesian Network model of this problem, and show that, while the witness's report does increase our probability of the cab being blue, it simultaneously decreases our estimate of their future accuracy (because blue cabs are less common). We presented this version of the problem to 131 participants, requiring them to update their estimates of both the probability the cab involved was blue, as well as the witness's accuracy, after they claim it was blue. We also required participants to explain their reasoning process and provided follow up questions to probe various aspects of their reasoning. While some participants responded normatively, the majority self-reported 'assuming' one of the probabilities was a certainty. Around a quarter assumed the cab was green, and thus the witness was wrong, decreasing their estimate of their accuracy. Another quarter assumed the witness was correct and actually increased their estimate of their accuracy, showing a circular logic similar to that seen in the confirmation bias/belief polarization literature. Around half of participants refused to make any change, with convergent evidence suggesting that these participants do not see the relevance of the witness's report to their accuracy before we know for certain whether they are correct or incorrect.

Keywords: causal Bayesian networks, second order uncertainty, propensity, uncertainty, confirmation bias

OPEN ACCESS

Edited by:
York Hagmayer,
University of Göttingen, Germany

Reviewed by:
Annelie Rothe-Wulf,
University of Freiburg, Germany
José C. Perales,
University of Granada, Spain

****Correspondence:***
Stephen H. Dewitt
dewitt.s.h@gmail.com

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

Received: *07 October 2019*
Accepted: *16 September 2020*
Published: *20 October 2020*

Citation:
Dewitt SH, Fenton NE, Liefgreen A and Lagnado DA (2020) Propensities and Second Order Uncertainty: A Modified Taxi Cab Problem. Front. Psychol. 11:503233. doi: 10.3389/fpsyg.2020.503233

INTRODUCTION

While causal Bayesian reasoning, and reasoning under uncertainty in general are major research programs within the judgment and decision-making literature, problems presented to participants have typically only studied this under first order uncertainty (also known as ‘risk’ in the economics literature). For example, the participant might be given a betting choice between a sure win of £25 or a 33% chance of £100 (e.g., Kahneman and Tversky, 1979). Here, while in the latter option it is uncertain whether we will get the £100, we can quantify this uncertainty precisely, and the problem thus yields simply to an expected utility calculation. But what if we did not know for certain what the chance of getting the £100 was? For example, suppose the probability was based on the outcome of some exotic asymmetrical die. Suppose also that we don’t understand the mechanics of the die, but we have observed 3 rolls, with only 1 leading to a win. While 33% might still be our best guess, with such a small sample size to estimate this, a substantial range of other probabilities are possible. How would this affect our decision over which bet to take? This uncertainty about our first order uncertainty is known as second order uncertainty (e.g., Kleiter, 2018), and we currently know little about how classic findings in the judgment and decision-making literature apply under such conditions.

Kahneman and Varey (1990) divided uncertainty along another dimension: internal uncertainty and external uncertainty (see also Juanchich et al., 2017). While internal uncertainty comes from our own ignorance about the world (e.g., the mechanics of the above die), external uncertainty comes from the propensity for an external causal system (such as the exotic die) to produce various outcomes or effects (e.g., a ‘win’). However, much we reduce our internal uncertainty about the mechanics of the die, we will only ever be able to predict what face will land up according to those propensities and never be able to guarantee a given outcome. This example illustrates an interaction between these two types of uncertainty which was not discussed in that paper. In this situation we have internal uncertainty about the propensity (external uncertainty) of the die to produce a ‘win.’ This is an extremely common situation – in fact, outside of contrived situations such as (standard) die rolls and coin flips, our estimates of the propensities for external causal systems to produce a given effect often comes with some internal (second order) uncertainty. Consider the propensity for a prisoner to reoffend or a patient to relapse or suffer complications. In each case the individual presumably has some true propensity (although this may fluctuate in a complex manner over time and context) but we only have limited information from which to estimate it. We are principally interested here in individuals’ ability to update propensity estimates in light of new information, i.e., update first order uncertainty estimates under conditions of second order uncertainty.

Approaches used to solve first order probability problems typically cannot be applied to second order problems. Knight (1921) gave the example of a picnic as a situation where first order techniques (e.g., expected utility calculations) were not workable. However, a true investment scenario, as opposed to the example we began the paper with is also insightful. When deciding whether to invest, one may use current and historical stock market figures, one’s feelings about and trust in the CEO and other bits of information such as a tip from an insider and other known markers of health. Under such conditions the probability of a positive return on investment cannot be reduced to a first order point estimate with no variance. Indeed, Mousavi and Gigerenzer (2014, 2017) have lamented the fact that while the vast majority of the experimental economics literature has aimed to study a higher order uncertainty problem (reasoning about business and economics), it has used experimental materials featuring only first order uncertainty. If we want to understand real world human reasoning outside of casino gambling, we must incorporate higher order uncertainty into the problems we use to study this.

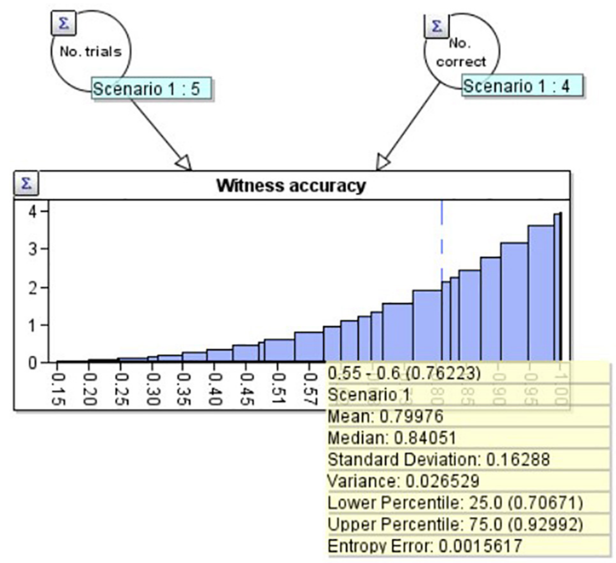
Similarly, while second order uncertainty has been written about in the context of causal Bayesian reasoning within the judgment and decision-making literature (e.g., Gigerenzer and Hoffrage, 1995; Welsh and Navarro, 2012; Kleiter, 2018), reasoning under these conditions has rarely been studied, and experiments aiming to study real world reasoning have also typically done this using problems with only first order uncertainty. For example, in the classic taxi cab problem (Tversky and Kahneman, 1974; Bar-Hillel, 1980), solvers are asked to reason about whether a cab involved in a hit and run accident was from the ‘blue’ company (as opposed to the ‘green’) in light of a population base rate (which suggests green cabs are more common) and an eye witness report (which claims a blue cab was involved). Solvers are told that the witness was tested for their ability to judge cab color, and that their ability was found to be 80%. A version of this can be seen below:

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city.
> You are given the following data: 90% of the cabs in the city are Green and 10% are Blue.
>
> A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each of the two colors 80% of the time and failed 20% of the time.
>
> What is the probability that the cab involved in the accident was Blue rather than Green?

In order to solve the problem, participants have to integrate the figure regarding the proportion of blue cabs in the city (not many: 10%) with the contradictory evidence of the witness’s claim that the cab was blue and their accuracy (quite good: 80%) to arrive at a final probability that the cab involved in the incident was blue. A major finding of the original paper was that many participants neglected the population base rate data entirely in their final estimate, simply giving the witness’s accuracy (80%) as their answer. Subsequent work has found that base rates more specific to the incident [e.g., in the area of the incident rather than in the city as a whole (Bar-Hillel, 1980)] and more causally related [e.g., where green cab drivers are known to get into more accidents, rather than just being more prevalent (e.g., Ajzen, 1977)] reduce base rate neglect.

In the similar medical diagnosis problem (e.g., Casscells et al., 1978; Gigerenzer and Hoffrage, 1995), solvers are asked to reason about whether a patient has cancer, given a population base rate (suggesting cancer is unlikely) combined with a positive test result. The solver is told that the false positive error rate of the mammogram test is 5%. In reality of course, there is likely to be some uncertainty around the probability estimates of both the witness's accuracy in the taxi cab problem and the false positive rate of the medical test. Our estimates therefore should look more like a distribution than a single point. While 80%/5% might provide the mean, or our best guess, there will also be some variance around this, due to our ignorance (internal uncertainty). The degree of variance depends upon the quality and amount of information we have available. These two examples prove useful in demonstrating this. While it may seem plausible that the mammogram machine has been tested a great many times, perhaps thousands of times, and thus, variance in our estimate might be very small, this seems less plausible for the single witness in the taxi cab problem, where time and resources would heavily limit the number of tests possible. Furthermore, it seems unlikely that the exact circumstances of the crash could be replicated for testing purposes, further increasing our uncertainty in the estimate. While we may therefore be justified in approximating the 5% false positive rate as a fixed-point estimate with no variance to simplify the problem, this is unlikely to be reasonable for the taxi-cab problem.

For example, suppose the witness has been tested 5 times, getting 4 correct. This produces a distribution with a classical statistical mean of 80% and a standard deviation of 16.3%. We created such a distribution in the 'AgenaRisk' Bayesian network program, which can be seen in **Figure 1**. We use a beta distribution (Kleiter, 2018) based upon the two nodes above it: No. trials (5) on the left, and No. correct (4) on the right. The mean and other statistics associated with the distribution can be seen in the yellow summary box.



0.55 - 0.6 (0.76223)
Scenario 1
Mean: 0.79976
Median: 0.84051
Standard Deviation: 0.16288
Variance: 0.026529
Lower Percentile: 25.0 (0.70671)
Upper Percentile: 75.0 (0.92992)
Entropy Error: 0.0015617

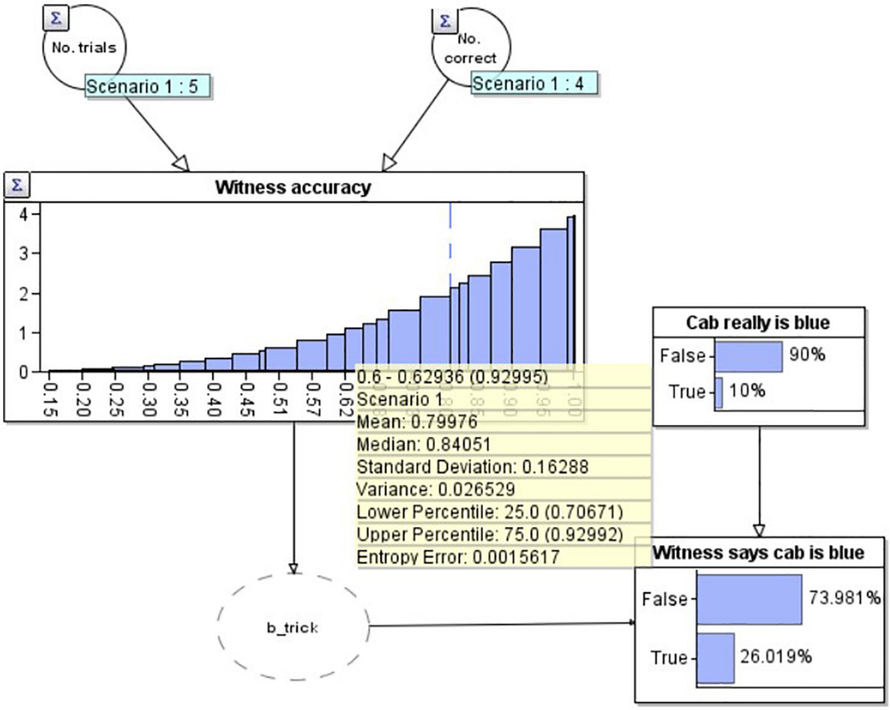
FIGURE 1 | A beta distribution with mean 80.0% and standard deviation 16.3% created using the Agenarisk software.

Now that we know the initial distribution of our estimate of the witness's accuracy, in order to model the full problem, we need to be able to update this distribution depending on whether the witness gets future reports correct or incorrect. We model this by expanding **Figure 1** into a larger Bayesian network (BN: **Figure 2**). A BN is a directed graph whose nodes represent uncertain variables, and where an arc (or arrow) between two nodes depicts a causal or influential relationship [see Fenton and Neil (2018) for full details of BN's]. In addition to the graph structure, each node has an associated probability table which defines the prior probability distribution for the associated variable, conditioned (where a node has parents) on its parent variables. When the state of a node is observed (e.g., the witness reports that the cab is blue) the known value is entered into the BN via an 'observation' and a propagation algorithm updates the probability distributions for all unobserved nodes. The 'Bayesian' in BN's is due to the use of Bayes' theorem in the underlying propagation algorithm.

In this diagram, our estimate for the witness's accuracy has been connected to a node ('Witness says cab is blue') depicting whether the witness reports that the current cab is blue. The 'b_trick' node is simply a pragmatic software requirement to convert the witness's accuracy distribution into a binary variable. The probability that the witness says the cab is blue is causally dependent upon both their accuracy, and the base rate, depicted in the above node 'Cab really is blue.' The current diagram depicts the situation before the witness makes their report. The best estimate that the cab is really blue at this point is just the base rate, 10%. Combining this and the witness's accuracy, the model predicts a 74.0% chance that the witness will report that the cab is green.

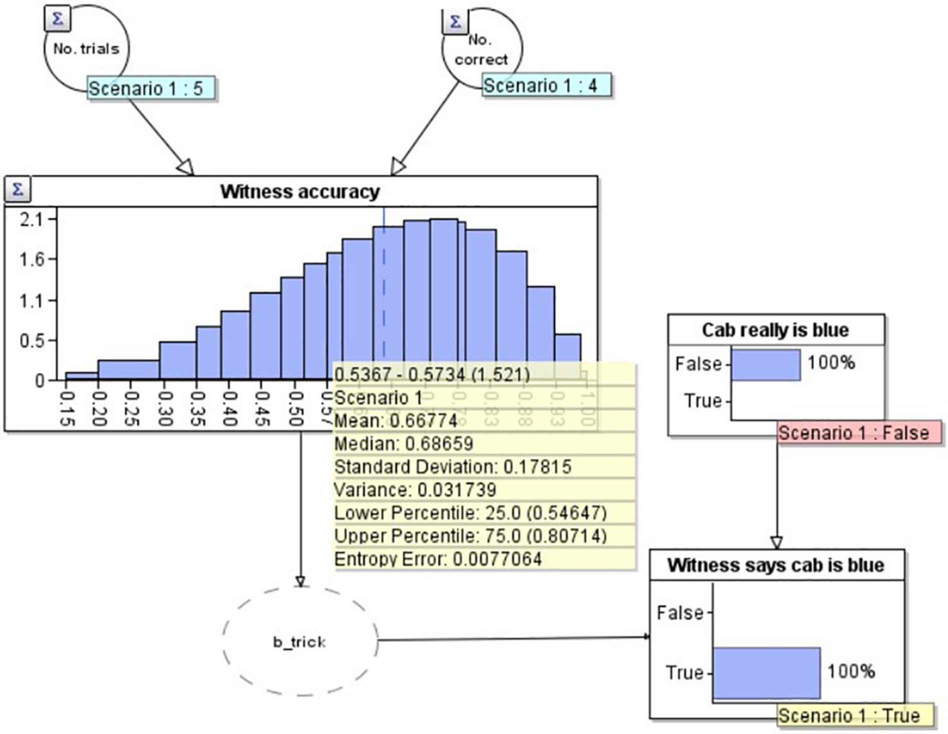
To demonstrate the workings of the model, in **Figure 3** we add two observations to the model. Firstly, in the lower right, we set an 'observation' on the 'Witness says cab is blue' node that the witness has said the cab is blue (note that this now has a yellow label saying 'True'). We have also set an observation on the 'Cab really is blue' node to make this 'False' i.e., as if we knew the cab really was green (and therefore the witness was incorrect). As would be expected in this scenario, our estimate of the witness's accuracy goes down to 66.7% (see yellow summary box mean), as we would expect given that they now have 4 correct out of 6 (4/6 = 0.666. . ..). Similarly, if the witness reports that the cab is blue, and we model that the cab really was blue (i.e., the witness is correct) by setting 'Cab really is blue' to 'True' (not depicted), the model provides an estimate of the witness's accuracy of 83.3%, equivalent to getting 5 out of 6 correct (5/6 = 0.833). Of present interest however, is how the witness's accuracy should change outside of these 'certain' bounds: when the witness reports that the cab is blue, but we don't know for certain if they are correct or not (**Figure 4**).

In **Figure 4** we have modeled the problem to include the witness's report that the cab was blue, but without knowing the truth for certain (no 'observation' on the 'Cab really is blue' node). Not only, as expected, has the probability that the cab is



```
0.6 - 0.62936 (0.92995)
Scenario 1
Mean: 0.79976
Median: 0.84051
Standard Deviation: 0.16288
Variance: 0.026529
Lower Percentile: 25.0 (0.70671)
Upper Percentile: 75.0 (0.92992)
Entropy Error: 0.0015617
```

FIGURE 2 | A Bayesian network depicting the modified taxi cab problem prior to the witness reporting the cab is blue.



```
0.5367 - 0.5734 (1.521)
Scenario 1
Mean: 0.66774
Median: 0.68659
Standard Deviation: 0.17815
Variance: 0.031739
Lower Percentile: 25.0 (0.54647)
Upper Percentile: 75.0 (0.80714)
Entropy Error: 0.0077064
```

FIGURE 3 | A Bayesian network model depicting the situation where the witness is incorrect about the cab being blue.

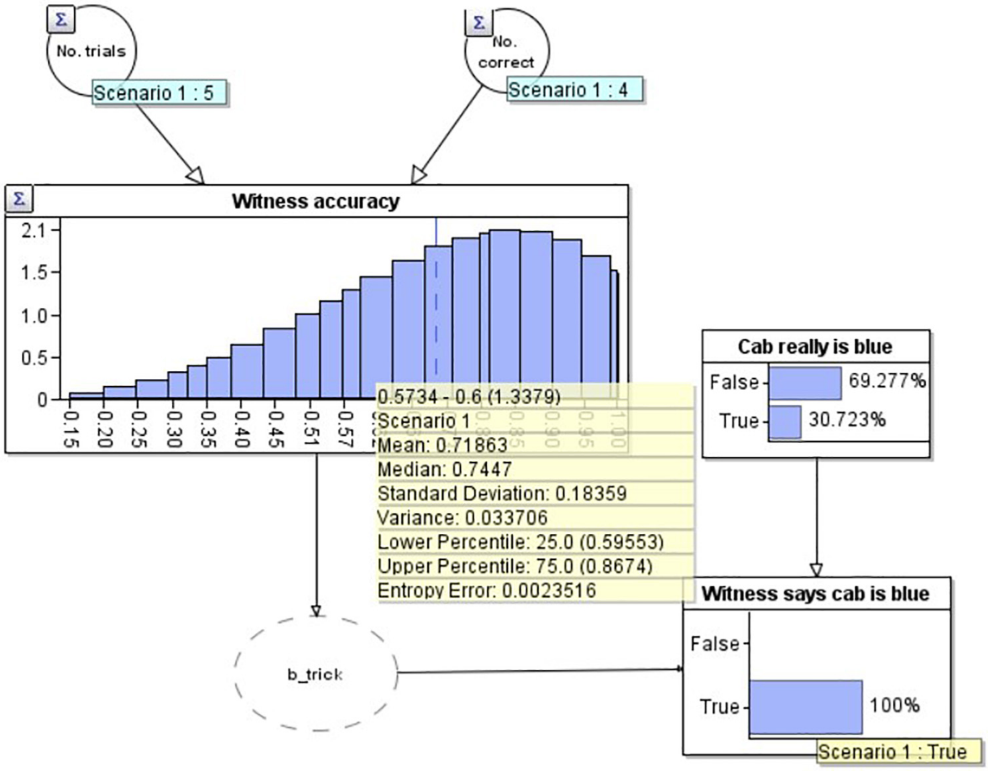


FIGURE 4 | A Bayesian network model depicting the situation where the witness has reported the cab is blue but we are uncertain if they are correct or incorrect.

blue increased (to 30.7% from 10.0%), but simultaneously, our estimate of the witness's accuracy has reduced, to 71.9%, below the initial estimate (80.0%, i.e., 4/5) but not as low as the estimate if we knew for certain the witness was incorrect (66.7%, i.e., 4/6). The reason for this reduction is that the witness has made a claim which goes against the only other evidence we have (the base rate, which suggests that the cab is green with considerable strength). If the witness had instead claimed that the cab was green, our estimate of their accuracy would increase (to 82.9%), again less than if we knew for certain they were correct (83.3%).

This addition of second order uncertainty therefore gives the problem a more dynamic character than the original problem. Furthermore, it cannot be solved with a simple application of Bayes' theorem, unlike the original taxi cab problem. It also has a potentially unintuitive dynamic: while we have enough trust in the witness to 'use' the information they provide as evidence that the cab was blue, we simultaneously reduce our trust in the witness's ability to make this very judgment in future. To keep things initially simple, as can be seen, we do not model the prior for the cab being blue as having second-order uncertainty. As will be discussed later, the version of the problem we use justifies a fixed estimate for this (we have complete knowledge), however, versions with second order uncertainty here may also be interesting.

Our primary aim is to examine participant responses to this novel problem and their ability to reason about causal relationships under second order uncertainty, and particularly through that unintuitive dynamic which is typical of such problems. Lacking the assistance of software like the above, the precise normative answer will not be achievable by our participants. For this reason, and because we believe such numerical precision is unlikely to characterize real world reasoning, we are not interested in participants' ability to do the mathematics, or the magnitude of their adjustments when they find out the witness reports the cab was blue. Instead we are interested only in the direction of their adjustments for the two main estimates (the witness' true accuracy level and the probability the cab is blue) and particularly whether they recognize that the witness's accuracy should be reduced. We will also request participants to explain their reason for their responses and provide several follow up questions to probe their representation and processing of the problem, in line with recent calls for more process-oriented work within this literature (Johnson and Tubau, 2015; McNair, 2015).

MATERIALS AND METHODS

Participants

One hundred and thirty-one participants (43.5% female), recruited from Prolific Academic (paid £9 per hour), took part in the study, with an average age of 27.8 (*SD* = 9.8). No participants were removed from the statistical analyses.

Design

All 131 participants saw the same version of the study. Participants were sub-divided in the analysis based upon their response to the key question of interest, and we used a range of numerical, and open and closed qualitative data to uncover the cognitive processes behind these different response types.

Materials and Procedure

All materials and data can be found in a public repository at https://osf.io/q68cu/. Participants were firstly presented with the information sheet and once clicking 'Next' to indicate their consent, were presented with the hit and run scenario. Participants were only able to move forward in the experiment, and could not go back and check previous pages. They were first told that a CCTV camera had made a 'partial read' on a taxi cab's license plate fleeing the scene, and that only 10 cabs matched: 1 belonging to the blue company, 9 to the green company (giving a first order probability with no second order uncertainty, assuming it is trusted). They were then asked to give a percentage estimate using a slider that the cab was blue based only on this information. On a new page, they were then told a witness had come forward, and were given information on the witness's accuracy. Participants were told the police had tested the witness five times, and the witness was correct four times. They were then asked to estimate the witness's true accuracy from this. To encourage participants not to see the initial accuracy value as fixed (i.e., overly subscribe to the law of small numbers), we included the following text to emphasize that we have only a limited estimate of their true accuracy.

However, we only have 5 trials to estimate this. It's possible they got lucky once or twice during the test. If we ran 100 trials we would have a more reliable estimate. Perhaps they would get 70 correct, or even 90.

Participants were then asked to use two sliders to give an estimate of the witness's 'true' accuracy (0–100%), and separately, provide a (0–100%) confidence that that estimate "would be the witness's true accuracy level if we ran a lot more trials." Only after providing these two prior estimates, participants were told on a new page that the witness had claimed the cab was blue.

Following this, participants were first asked to update their estimate that the cab was blue and then on a separate page, update their estimate of the witness's accuracy. In both cases key information was re-summarized. Instead of being asked to give a numerical value at this point, participants indicated on a sliding scale (**Figure 5**) whether they wanted to stick with their original value or increase their estimate. Participants were forced to answer all questions in the survey and were not able to proceed with the experiment if they simply left the slider in place. If they wished to make no change and keep their original estimate they first had to move the slider to activate it, then move it back to the center.

The degree to which participants moved the slider was not of importance, and was only included to allow participants to express themselves and to reduce the chance of participants who wanted to make a very small change choosing to make no change. This approach was used to discourage participants from attempting a mathematical treatment of the problem, which we strongly believe cannot be the way people solve real life problems of this type. Instead, we wanted to capture intuitive feelings of whether the two variables both go up, both go down, stay the same, or (as predicted by the normative model) the probability of the cab being blue goes up, while the accuracy of the witness goes down. It is at this coarser level at which participants responses were judged. For both estimates, on the same page, participants were asked to explain their reasoning in an open text box.

After making posterior estimates, participants were asked in a multiple-choice format whether, when reasoning through the problem they had (A) Assumed the cab was green, (B) Assumed the witness was correct or (C) Neither/Other. The order of these options was randomized.

Participants were finally told on a separate page that after the investigation had concluded it turned out the cab really was green and so we now know the witness was incorrect this time. Participants were then asked again whether they wished to adjust the witness's accuracy using the same slider and were again provided with an open text box to explain their reasoning.

RESULTS

Manipulation Checks/Priors

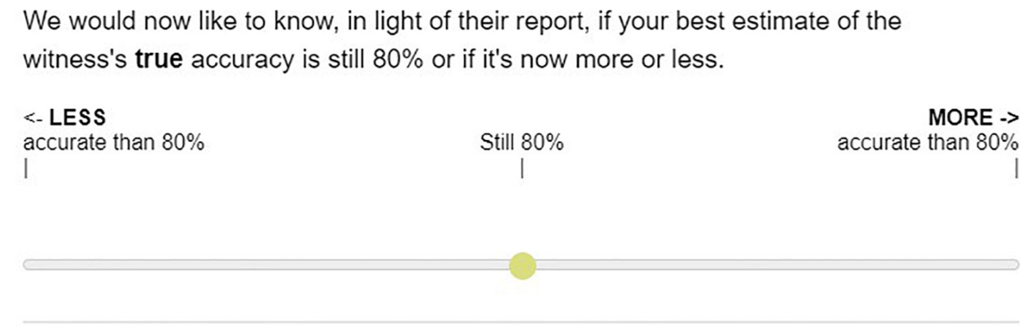
After being provided with the prior for the cab being blue, participants were asked to indicate on a sliding scale the probability the cab was blue and 82.4% chose 9, 10, or 11%, suggesting a high level of 'acceptance' of the prior figure (Mean = 15.6%, *SD* = 16.2%). Participants were also asked to do the same for the witness's accuracy, after being given the figures on the court's testing of them and 45.8% chose 79, 80, or 81% (Mean = 73.5%, *SD* = 15.5%). The distribution of responses for both can be seen in **Figure 6**. Participants were also asked to express their confidence that this figure represented the witness's true accuracy, which produced a mean of 68.6% (*SD* = 20.7%).

Posteriors

Once the witness reports that the cab was blue, participants were firstly asked to adjust their estimate that the cab was blue, and then the witness's accuracy. Out of all participants, 64.9% increased the probability that the cab was blue. Our primary interest, however, was what change they made to their estimate of the witness's accuracy. Only 21.4% reduced this, while 55.7% made no change, and 22.9% increased it. In the following we analyze these three sub-groups according to their responses to a range of questions to attempt to understand their cognitive processes. Figures to support these analyses can be seen in **Table 1** and will be referred to throughout.

Statistical Comparisons

We proceed in the following from top to bottom. In the second row of **Table 1** [(Mean) Cab blue %] we can see the average estimate that the cab was blue made by each of the three response types before the witness's report. A univariate analysis was run to test the effect of 'Response type' on '(Mean) Cab blue %'



We would now like to know, in light of their report, if your best estimate of the witness's **true** accuracy is still 80% or if it's now more or less.

<- **LESS**
accurate than 80%

Still 80%

MORE ->
accurate than 80%

FIGURE 5 | Image of the slider mechanism used for participants to adjust their estimate for the witness.

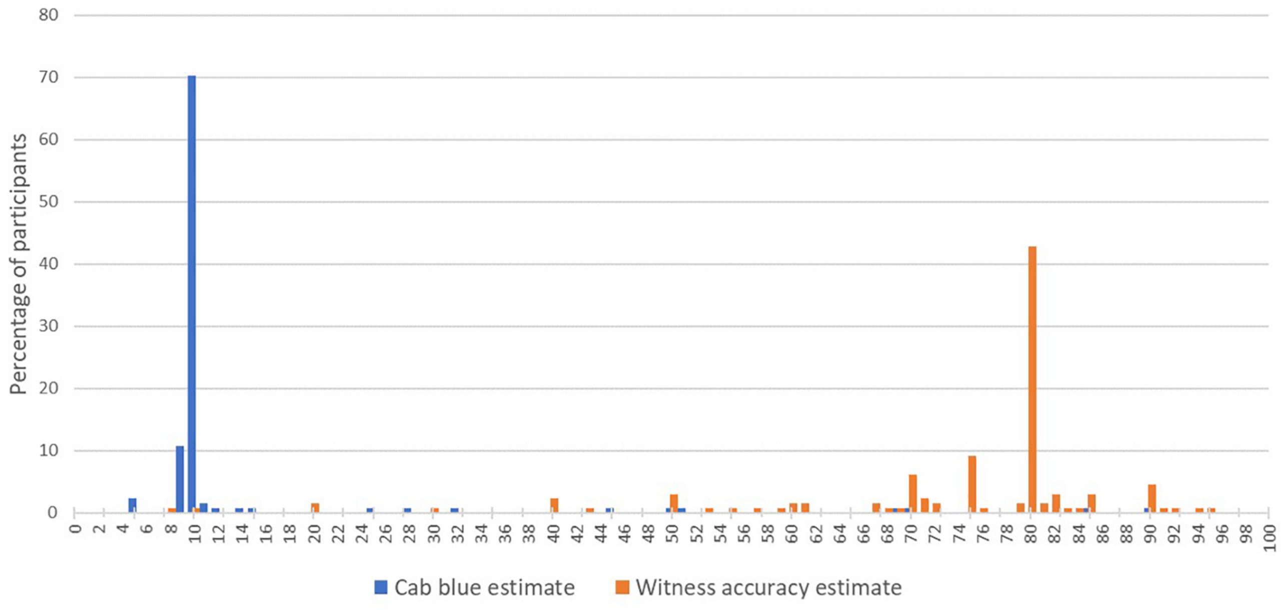


FIGURE 6 | A histogram representing participant prior estimates of the probability the cab is blue (blue) and witness accuracy (orange).

[F(2,128) = 2.6, p = 0.08]. Pairwise comparisons compared 'No Change' to 'Reduce' (p = 0.04) and 'No Change' to 'Increase' (p = 0.70) and 'Reduce' to 'Increase' (p = 0.04). Univariate analyses were also run to test the effect of 'Response type' on the third row, their estimate of the witness's accuracy before the witness's report [(Mean) Witness accuracy] [F(2,128) = 0.6, p = 0.55] and separately on row four, their confidence that this estimate was equal to the witness's true accuracy, [(Mean) Witness confidence] [F(2,128) = 0.13, p = 0.88].

Moving to row five, only a single individual reduced their estimate of the cab being blue after the witness' report. The remainder either made no change, or increased their estimate. A binary logistic regression was run to test the effect of 'Response type' on the proportion of individuals increasing their estimate of the cab being blue [Wald X^2 (2) = 6.6, p = 0.04]. Pairwise comparisons were then run to compare 'No Change' and 'Increase' [Wald X^2 (1) = 6.3, p = 0.01], 'No Change' and 'Reduce' [Wald X^2 (1) = 1.1, p = 0.29], and 'Reduce' and 'Increase' [Wald X^2 (1) = 1.8, p = 0.18].

We now move to rows six, seven, and eight, representing the proportion of each response type who chose either 'Assumed the witness was correct,' 'Assumed the cab was green,' or 'Neither/other' when faced with this question. Binary logistic regressions were run to test the effect of 'Response type' on assuming the witness was correct [Wald X^2 (2) = 16.2, $p < 0.001$], on assuming the cab was green [Wald X^2 (2) = 14.9, $p < 0.001$] and on 'Neither/other' [Wald X^2 (2) = 1.2, p = 0.55]. Examining the assumption that the witness was correct, pairwise comparisons were run to compare 'No Change' to 'Increase' [Wald X^2 (1) = 3.5, p = 0.06], 'No Change' to 'Reduce' [Wald X^2 (1) = 9.3, p = 0.002], and 'Reduce' to 'Increase' [Wald X^2 (1) = 15.5, $p < 0.001$]. Examining the assumption that the cab was green, pairwise comparisons were run to compare 'No Change' to 'Increase' [Wald X^2 (1) = 2.4, p = 0.13], 'No Change' to 'Reduce'

TABLE 1 | Participant responses to a range of questions sub-divided by their initial response to altering the witness's accuracy.

	Reduce	No change	Increase
Total *N*	28 (21.4%)	73 (55.7%)	30 (22.9%)
	Cab/witness statistics provided		
(Mean) Cab blue estimate	9.5% (0.3%)	16.8% (2.0%)	18.2% (3.6%)
(Mean) Witness accuracy estimate	76.0% (1.9%)	73.3% (2.1%)	71.6% (2.6%)
(Mean) Confidence	67.1% (2.9%)	68.8% (2.7%)	69.9% (3.3%)
	Witness reports cab is blue		
(Proportion) Increasing estimate cab blue	67.9% (9.0%)	56.2% (5.8%)	83.3% (6.9%)
(Proportion self-reported assuming):			
Witness correct	25.0% (8.3%)	60.3% (5.8%)	80.0% (7.4%)
Cab green	46.4% (9.6%)	15.1% (4.2%)	3.3% (3.3%)
Neither/other	28.6% (8.7%)	24.7% (5.1%)	16.7% (6.9%)
	Told cab actually green		
(Proportion) Reducing witness accuracy	78.6% (7.9%)	52.1% (5.9%)	73.3% (8.2%)

In brackets on the left it is indicated whether the figure provides the mean response, or the proportion providing a particular response (in the case of non-continuous outcomes). Questions are in chronological order and the delivery of key pieces of information is indicated. Standard errors for estimates are included in brackets.

[Wald X^2 (1) = 10.0, p = 0.002] and 'Reduce' and to 'Increase' [Wald X^2 (1) = 8.8, p = 0.003].

Moving to the final row, where individuals were asked to update their estimate of the witness's accuracy again after being told that subsequent investigations had found the cab really was green, only seven individuals increased their estimate of the witness's accuracy. All others either reduced or made no change to their estimate. A binary logistic regression was run to test the effect of 'Response type' on the proportion reducing their estimate [Wald X^2 (2) = 7.7, p = 0.02]. Pairwise comparisons were run to compare 'No Change' to 'Increase' [Wald X^2 (1) = 3.8, p = 0.05], 'No Change' to 'Reduce' [Wald X^2 (1) = 5.5, p = 0.02] and 'Reduce' to 'Increase' [Wald X^2 (1) = 0.22, p = 0.64].

Qualitative Data

Participants were asked to explain their reasoning after providing their posterior change estimate for the Witness's accuracy. These were coded blind to response type by the first author. Four major codes were identified, but around half of all responses were also coded as 'Unclassified' where an understanding of the participants' response could not be confidently attained. The first author gave their codebook containing these five codes (**Table 2**) to the third author. The third author then assigned these codes, blind to both response type and to the first author's assignments. Inter-rater agreement was 78.6%, with disagreements generally being whether a response should be 'unclassified' or not. For the discrepant responses (28 total), if one coder had chosen 'unclassified' we assigned this code in order to be conservative – 22/28 of these were therefore classified that way. The remaining six were resolved through discussion. The proportion of each response type assigned each code post-agreement can be seen below in **Table 2**. Among responses that could be classified, one modal code stands out for each, however, for 'No Change,' a substantial amount were also coded as 'Witness probably correct,' similar to the 'Increase' responders. These will be discussed below.

Increase

The modal code assigned among 'Increase' responders was 'Witness probably correct.' This was assigned where the participant indicated that they thought the witness was likely to be correct or showed confidence in the witness. A selection of these responses can be seen in **Table 3**.

No Change

The modal code among 'No Change' responders (30.8%) was 'Irrelevant.' This was assigned where the participant stated that the report by the witness has no bearing on their accuracy level. A selection of these responses can be seen in **Table 4**.

When told at the end of the experiment that the cab really was green, and the witness was incorrect, we can see that half of 'No change' participants still made no change to their estimate of the witness's accuracy. A selection of these participants' explanations of those responses can be seen in **Table 5**.

Reduce

The most prominent code among 'Reduce' responders was 'Witness probably incorrect.' This was assigned when the participants stated that the witness was probably incorrect on this occasion, or expressed low confidence in them. Some of these also referenced the low base rate for blue cabs. A selection of these responses can be seen in **Table 6**.

DISCUSSION

In this paper we aimed to examine responses to a modified version of the classic taxi cab problem including second order uncertainty. Through mixed methods, we aimed to uncover participants' approaches to handling the new dynamics introduced in the modified version. Of principal interest was how participants altered their estimate of the witness's accuracy after the witness reported that the cab was blue. We found that around half made no change, with around a quarter

TABLE 2 | Percentages of each response type assigned each code type (modal code excluding 'unclassified' is highlighted for each response).

	Witness probably correct	Irrelevant	Witness probably Incorrect	Requires Certainty	Unclassified
Increase	33.3	3.3	6.7	–	56.7
No change	17.8	32.9	8.2	2.7	38.4
Reduce	3.6	–	50.0	–	46.4

each reducing/increasing their estimate of the witness' accuracy. Through convergent evidence, combining quantitative and qualitative responses we present below a general picture of the cognitive processes involved in each response type, however, we do not assume each represents a single coherent population (multiple cognitive processes may lead to the same response) and in some cases suggest this might be the case.

Increase

'Increase' responders appear to be the most homogeneous of the three response types, with 80% self-reporting as having assumed the witness was correct, and, outside of the large proportion 'unclassified,' the majority of their open text data being coded as 'Witness probably correct.' Interestingly, none of these responders explicitly say they are ignoring the base rate data, or that they trust the witness more than the base rate data. They generally just only refer to the witness data in their responses, expressing confidence in them or belief they are correct, and typically not mentioning the base rate data at all. These responders cognitive process may therefore represent a form of base rate neglect (e.g., Tversky and Kahneman, 1974; Bar-Hillel, 1980). However, as can be seen in **Figure 6** and **Table 1**, these participants certainly saw the base rate data as relevant before the witness made their report, suggesting a simple disregard for the relevance of that information is not a good explanation. However, Bar-Hillel (1980) proposed a 'dominance' theory of base rate neglect in such problems, where the piece of information seen as least relevant would be entirely disregarded, presumably for reasons of computational simplification. It is possible that our participants, once the witness report is provided, find the prospect of integrating these two figures too daunting. From there, finding the witness report more compelling than the base rate data for whatever reason, they may disregard the base rate, leading to an 80% estimate that the cab is blue based solely on the witness's accuracy. However, this cannot provide a full explanation of the present results. Even if these participants do believe there is an 80% chance that the cab is blue, how does this justify increasing their estimate of the witness's accuracy?

A similar response was also detected in other papers by the authors on reasoning with propensities (Dewitt et al., 2018, 2020). In the scenario presented in both those papers, two nations are testing their missile detonation capabilities. Nation *X* has so far had only 1 success out of 6 attempts while *Y* has had 4 successes out of 6. Another missile then successfully detonates on the border between the two nations but we're unable to detect the source. The key question, instead of who launched this missile (equivalent to whether the cab is blue), is what the new proficiency estimate for each nation is (equivalent to updating the witness's accuracy). In both papers we found a similar approach

TABLE 3 | A selection of 'Increase' responders open-text explanations of their reasoning assigned the code 'Witness probably correct'.

"Because he got another car right so it is 5/6."
". . . Now with 80% accuracy of the witness, stating that it is from the blue one, I am sure that the cab is from the blue company."
"The witness has said they saw a blue cab."
"I believe the witness would have seen correctly even under pressure."
"They did do the trial 5 times and out of the 5 times they got it correct 4 times, had this been lower then I would have questioned the accuracy but 4 out of 5 is quite good."

TABLE 4 | A selection of 'No Change' responders open text explanations of their reasoning assigned the code 'Irrelevant'.

"Still 90% because the facts are still the same."
"I don't think the probability of what they saw regarding colors will affect the accuracy of their statement."
"The report doesn't change their estimated accuracy."
"For me nothing changed because we have no new viable information."
"Because the probability of it being a blue car and then the witness identifying it as a blue car are separate, so even if it's a low probability, it wouldn't affect their perception unless they were told beforehand that [it was] low probability."
"The result of the test (blue or green) doesn't change the level of accuracy of the witness."
"They still got 4/5 trials right, so I'm still confident in them."

TABLE 5 | A selection of 'No Change' responders open text explanations of their reasoning after being told the witness was incorrect and still making no change to their estimate of the witness's accuracy.

"The witness managed to get the correct color 4/5 based on the test. 1/5 times the witness fails and this was one of the situations where they failed."
"I don't feel that I can judge their accuracy based on this as this result could have been in the 20%"
"No remains 80%. The 20% percent would be them getting the color wrong."
"I believe their accuracy is still not in question, they still had a 1 in 5 chance to get it wrong."
"The previous test measured that the witness had a 4/5 chance to get the color correct. The accuracy still stands."
"It fits 20% of not getting the right color."
"There was still 20% chance he was wrong."
"He has 4/5 so the car could be the 1/5."

TABLE 6 | A selection of 'Reduce' responders open text explanations of their reasoning assigned the code 'Witness probably incorrect.'

"Based on the potential cab colors, it's more likely than not that the cab was green, so I'm slightly more inclined to doubt the witness."
"Very unlikely that it was a blue cab, since only one out of 10 plates were blue."
"Because people can think they saw a thing and can be another completely different."
"From a statistical point of view it is likely that the witness was wrong."
"It was considerably less likely to be blue than green; this coupled with the one incorrect trial result, makes me less confident that the witness is correct. However, they still actually could be."
"If was dark, how he/she can know whether car blue or green?"
"I think the accuracy of the witness became less than 80% because the probability that was a green cab is higher than her accuracy."
"It's hard to identify the color of a moving car at night, besides blue and green at high speed are easy to mistook for each other."

to the present paper, where, when updating proficiencies, 1/3 of participants increased their estimate of *Y*'s proficiency, making no change to *X*. Because *X* and *Y* represent exhaustive and exclusive causes of the latest explosion, unlike in the present scenario, we were able to infer that they have treated *Y* launching the latest missile as a certainty (i.e., they have 'given' the whole responsibility for the detonation to *Y*, and none to *X*). In Dewitt et al. (2018) we also saw similar open text reasoning, with the majority of these participants simply stating that they believed *Y* was probably responsible. We labeled these participants as 'categorical' responders as we found that while they rated the probability that *Y* was responsible as 77.7% on average, they all treated it as a 100% certainty when updating their estimate. This treatment of a probabilistic variable (e.g., 80% estimate that the cab is blue based on the witness's report) as a certainty may also be occurring in the present experiment. Evidence for this comes from the vast majority of 'Increase' responders who self-report as having 'assumed' the witness was correct.

Importantly, in both scenarios, this response has a circularity to its logic. In the missiles experiment, *Y* is assumed to have launched the missile based upon their previous success with missiles. Their success with missiles is then updated based upon the assumption that *Y* launched the missile (which, to reiterate the circularity, is based upon their past success with missiles). Similarly, in the present scenario, the only evidence that participants have that the witness might be correct this time is their previous accuracy. But just like in the missiles scenario, this is the very thing we want them to update. So, when this approach is adopted, it appears that once a witness gets to a certain level of trust, they will not only be assumed to be correct based only on this historical accuracy, but also, based on that assumption, they will be seen as even more accurate afterward, even when the only other evidence available actually strongly suggests they are incorrect. This has the same circularity as has been observed in confirmation bias and belief polarization literature (Lord et al., 1979; Plous, 1991; Nickerson, 1998; Cook, 2016; Fryer et al., 2019).

The treatment of probabilistic variables as categorical variables has also been previously reported under the names of 'as if' reasoning (Gettys et al., 1973) and 'digitization' (Johnson et al., 2018). Both sets of authors have found that in multi-step reasoning, where the output of one probabilistic calculation is used in a second calculation, the first output is often digitized (or, turned into categorical form) for the second calculation. For example, if one has to calculate the chance of rain, and then use that probability to estimate the chance that a party will be canceled, they will treat the chance of rain in that second calculation as either 0 or 1. In our missile launching scenario, this 'multi step' explanation for categorization was considered plausible, as the categorical response involved multiple steps (one first had to normalize a 66.6:16.6 ratio to get to the 80:20 probability of who launched the missile before using this latter value to update propensities). However, the present problem does not involve multiple steps as the participants are directly provided with the probabilities of, e.g., the cab being green and the witness being correct. While these are admittedly estimated from frequencies (9/10 and 4/5) this is not a true 'first' calculation in the sense meant by those authors (e.g., if participants had to first multiply two figures together to get the witness's accuracy). Therefore, despite this problem not fitting the 'multi-step' format, we still saw large numbers of participants taking what appears to be a similar categorical approach. This may suggest that this phenomenon of digitization or categorization is a more general strategy to simplify a difficult problem (with multiple steps being just one source of difficulty). The current scenario unfortunately presents a situation (with two diagnostically opposite pieces of data) where such a strategy is at its most inappropriate (unlike, e.g., if one of the figures was close to 50:50), and so it is interesting that we still see such a strategy employed. It may be valuable to determine in future work if individuals are sensitive to the 'appropriateness' of this strategy when choosing to employ it, by varying the figures in the problem. It should also be noted that, even if participants are aware that it is not an ideal approach to the scenario, they may feel that they lack an alternative approach. Indeed, in Dewitt et al. (2020) we found that 1/3 of categorical responders endorsed the statement 'I approximated that *Y* was entirely responsible for the launch in order to make the problem simpler but know this is not strictly accurate,' suggesting some awareness that their approach was not fully normative.

Reduce

'Reduce' responders appear to be more mixed than 'Increase' responders in their choices on the 'assumption' question. Around half self-report as assuming the cab is green, but around a quarter actually report they assumed the witness was correct, and another quarter report 'Neither/other.' However, almost all of those whose open text responses could be classified were coded as 'Witness probably incorrect.' Unlike 'Increase' responders, many of these did cite the base rate as a reason for this belief, saying either that the cab was very unlikely to be blue, or very likely to be green. Others stated low confidence or disbelief in the witness's ability to make the judgment.

We think it is possible that there are two sub response types here. First, would be the mirror image of the 'Increase' responders, who may be committing 'base rate conservatism,' neglecting the relevance or value of the witness's claim, and entirely focusing on the base rate and treating that as a certainty. This would correspond to those 46.4% who self-reported as assuming the cab was green.

Second would be those who dealt with the problem normatively/probabilistically, integrating both variables together (even if not fully mathematically). As the base rate is stronger than the witness's report (90% vs. 80%), this leads to the conclusion that the witness is more likely to be incorrect than correct. There is no equivalent process to this that would lead to 'increase,' which is another reason to suppose a single cognitive process for that response. This process may correspond to those 28.6% who selected 'Neither/other' (i.e., didn't 'assume' either way). These participants may have, rather than neglecting one piece of information or the other (either the base rate or the witness information), have integrated both, concluding that the witness is more likely to be incorrect but still maintaining a probabilistic representation of the problem, rather than collapsing into assumption-based thinking.

No Change

Participants who made no change are also less obviously homogenous than the 'Increase' responders. Interestingly, quite a large proportion state that they assumed the witness was correct, despite not increasing their estimate of the witness's accuracy. Similar to 'Reduce' responders, around a quarter self-reported assuming 'Neither/other.'

While the main qualitative code for 'No change' responders was 'Irrelevant,' and only a few were coded as 'Requires uncertainty,' it is still very possible that uncertainty is a major reason for this response. These participants tended to be quite unforthcoming in their reason for why they think the new information is irrelevant, simply stating that they don't see the connection. The reason for the new information being irrelevant could therefore very well be that we don't yet know whether the cab really is green or blue – they may consider uncertain information irrelevant for updating the witness's accuracy. This would fit with the 'missiles' experiments mentioned above (Dewitt et al., 2018, 2020) where 'No change' was also a dominant response (about 1/3 both samples). In Dewitt et al. (2020), when asked to pick from a set of statements as to which most closely matched their reasoning, around two thirds of these responders chose 'The evidence states it's uncertain who launched the successful missile so you cannot change the proficiencies based on uncertainty.'

Another reason to think that uncertainty may be an important underlying reason for the 'No change' response is that half of these participants, when told that the cab really was green at the end of the experiment (and therefore that the witness was incorrect), reduced their estimate of the witness's accuracy. This suggests that they do see the connection between the witness's report that the cab was blue and their accuracy, but only once they know for certain that that report really is wrong. Indeed, while under uncertainty, participants may prefer to err on the side of avoiding updating incorrectly (Anderson, 2003).

However, a much larger percentage of 'No change' responders than either 'Reduce' or 'Increase' continue to make no change even when told that the witness was incorrect. A selection of responses which may indicate these participants' thought processes can be seen in **Table 5**. There is a strong theme here of these participants seeing the latest failure of the witness as 'fitting within' the original accuracy estimate of 80%. In some cases, they seem to suggest that this could be 'the one' they got wrong (out of 5), which seems obviously incorrect given that previous information told them they already got one wrong during the tests. It is difficult to tell therefore whether this represents a simple misunderstanding of that, or whether there is a deeper and more interesting process occurring. Indeed, another interpretation is that these participants see one additional data point (even if that data point is now certain) as not enough to change a propensity based on five data points, when that propensity allows for some failure (i.e., the 20%). These participants may be seeing this latest claim as the first data point in another 'run' of 5, and while this first one failed, the next four may be successes, matching the original '80%.' If true, there therefore seems here to be an over-sanctification of the original run of data, and furthermore, a similar tendency to 'wait for more data before updating' as with the single no change response.

CONCLUSION

Overall, these findings seem to represent a general unwillingness or inability by at least 3/4 of our participants to deal with the problem probabilistically when answering second order questions, either converting those variables into categorical form (those 'Increase' and 'Reduce' responders who 'assumed' either way), or withholding judgment until they are certain ('No change' responders). This appears to represent a major departure from a Bayesian treatment of the problem, where any information about the state of one variable (i.e., the witness's report), even if probabilistic, can be used to update our estimates for other causally related variables (i.e., the witness's accuracy).

Generally therefore, in studying responses to this modified taxi cab problem, we have corroborated findings in previous work. The present work and the missiles work probe participants reasoning in different ways, and the problems have slightly different dynamics, and yet both point toward a substantial majority of participants adopting a categorical representation when updating propensities. Two approaches seem to stem from this. Some participants refuse to update entirely, until the state of the event is known. Other participants seem to convert the event into a certainty one way or the other, and update propensities based upon that assumption. While the issue with the former approach is to make no use of valuable information, the latter may be more damaging. We have already spoken about the circularity of the 'assume witness correct' approach. Before even knowing whether the witness is correct, the mere fact that the witness has shown themselves to be fairly reliable in the past, seems to lead these individuals to increase their trust in them following their claim, as if they knew the witness was correct this time. This suggests that once a person or system reaches a certain level of trust, they may be able to make claims, and even without the truth of these being determined, are not only trusted in the individual situation (which may be reasonable), but even have trust in them increased for the future on the assumption they probably were correct this time. This is perhaps all the more troubling given that the previous accuracy estimate in this experiment was based upon a very small number of tests of the witness. With such a small number of trials, it is highly possible that the witness just got lucky on a couple of occasions. With only two options to guess from, it is quite possible that they are at chance level for judging cab colors under the given conditions. Therefore, in combination, this seems to suggest that if an individual gets lucky with a few accurate claims early on, and they pass the 'safe to assume they are correct' threshold, their early luck can take on a self-reinforcing dynamic, where trust in them is further enhanced even without further verification of their claims. This dynamic was also discussed in Dewitt et al. (2020) in the context of prejudice toward an individual or group of producing some negative outcome. Indeed, it may be that this 'assumption' approach lends any situation where we are estimating a propensity (whether for a 'good' or 'bad' outcome) a positive feedback dynamic similar to that seen in confirmation bias/belief polarization literature.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: https://osf.io/q68cu/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UCL Department of Experimental Psychology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

This work builds upon previous work by the current authors. SD conceived of applying those previous ideas to the present problem, developed, ran and analyzed the experiment, and was the primary author. Ideas were discussed and refined with DL throughout the entire process, including the analysis of results and DL provided thorough feedback on various drafts of the manuscript. NF provided the causal Bayesian network model and technical advice throughout including feedback on the manuscript. AL provided second coding for the qualitative data and feedback on the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *J. Pers. Soc. Psychol.* 35, 303–314. doi: 10.1037/0022-3514.35.5.303

Anderson, C. J. (2003). The psychology of doing nothing: forms of decision avoidance result from reason and emotion. *Psychol. Bull.* 129, 139–167. doi: 10.1037/0033-2909.129.1.139

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychol.* 44, 211–233. doi: 10.1016/0001-6918(80)90046-3

Casscells, W., Schoenberger, A., and Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *N. Engl. J. Med.* 299, 999–1001. doi: 10.1056/nejm197811022991808

Cook, J. (2016). Rational irrationality:modeling climate change belief polarization using bayesian networks. *Top. Cogn. Sci.* 8, 160–179. doi: 10.1111/tops.12186

Dewitt, S. H., Adler, N., Fenton, N. E., and Lagnado, D. A. (2020). Categorical Updating in a Bayesian Propensity Problem. Manuscript submitted for publication. Experimental Psychology, University College London.

Dewitt, S. H., Lagnado, D. A., and Fenton, N. (2018). "Updating prior beliefs based on ambiguous evidence," in *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, United States.

Fenton, N., and Neil, M. (2018). *Risk Assessment and Decision Analysis with Bayesian Networks*, 2nd Edn. Boca Raton, FL: Chapman and Hall / CRC Press.

Fryer, R. G., Harms, P., and Jackson, M. O. (2019). Updating beliefs when evidence is open to interpretation: implications for bias and polarization. *J. Eur. Econ. Assoc.* 17, 1470–1501. doi: 10.1093/jeea/jvy025

Gettys, C. F., Kelly, C., and Peterson, C. R. (1973). The best guess hypothesis in multistage inference. *Organ. Behav. Hum. Perform.* 10, 364–373. doi: 10.1016/0030-5073(73)90024-x

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295x.102.4.684

Johnson, E. D., and Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Front. Psychol.* 6:938. doi: 10.3389/fpsyg.2015.00938

Johnson, S., Merchant, T., and Keil, F. (2018). Belief digitization: do we treat uncertainty as probabilities or as bits? *SSRN Electron. J.* 1932, 1–20.

Juanchich, M., Gourdon-Kanhukamwe, A., and Sirota, M. (2017). I am uncertain" vs "it is uncertain". How linguistic markers of the uncertainty source affect uncertainty communication. *Judg. Decis. Mak.* 12, 445–465.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–292. doi: 10.2307/1914185

Kahneman, D., and Varey, C. A. (1990). Propensities and counterfactuals: the loser that almost won. *J. Pers. Soc. Psychol.* 2, 1101–1110. doi: 10.1037/0022-3514.59.6.1101

Kleiter, G. D. (2018). Imprecise uncertain reasoning: a distributional approach. *Front. Psychol.* 9:2051. doi: 10.3389/fpsyg.2018.02051

Knight, F. H. (1921). *Risk, Uncertainty, and Profit*. Boston, MA: Houghton Mifflin Company.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* 37, 2098–2109. doi: 10.1037/0022-3514.37.11.2098

McNair, S. J. (2015). Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. *Front. Psychol.* 6:97. doi: 10.3389/fpsyg.2015.00097

Mousavi, S., and Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. *J. Bus. Res.* 67, 1671–1678. doi: 10.1016/j.jbusres.2014.02.013

Mousavi, S., and Gigerenzer, G. (2017). Heuristics are tools for uncertainty. *Homo Oeconomicus* 34, 361–379. doi: 10.1007/s41412-017-0058-z

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220. doi: 10.1037/1089-2680.2.2.175

Plous, S. (1991). Biases in the assimilation of technological breakdowns: do accidents make us safer? *J. Appl. Soc. Psychol.* 21, 1058–1082. doi: 10.1111/j.1559-1816.1991.tb00459.x

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

Welsh, M. B., and Navarro, D. J. (2012). Seeing is believing: priors, trust, and base rate neglect. *Organ. Behav. Hum. Decis. Process.* 119, 1–14. doi: 10.1016/j.obhdp.2012.04.001

Conflict of Interest: 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.

PERSPECTIVE
published: 21 October 2020
doi: 10.3389/fpsyg.2020.527114

Cognitive Structures of Space-Time

Camilo Miguel Signorelli [1,2], Selma Dündar-Coecke [3], Vincent Wang [1] and Bob Coecke [1]*

[1] Department of Computer Science, University of Oxford, Oxford, United Kingdom, [2] Cognitive Neuroimaging Unit, INSERM U992, NeuroSpin, Gif-sur-Yvette, France, [3] Center for Educational Neuroscience/Department of Psychology and Human Development, University College London, London, United Kingdom

In physics, the analysis of the space representing states of physical systems often takes the form of a layer-cake of increasingly rich structure. In this paper, we propose an analogous hierarchy in the cognition of spacetime. Firstly, we explore the interplay between the objective physical properties of space-time and the subjective compositional modes of relational representations within the reasoner. Secondly, we discuss the compositional structure within and between layers. The existing evidence in the available literature is reviewed to end with some testable consequences of our proposal at the brain and behavioral level.

Keywords: causal cognition, causal structure, causality, space-time, compositionality

OPEN ACCESS

Edited by:
Snehlata Jaswal,
Chaudhary Charan Singh University,
India

Reviewed by:
Naotsugu Tsuchiya,
Monash University, Australia
Roberto Bottini,
University of Trento, Italy

****Correspondence:***
Camilo Miguel Signorelli
cam.signorelli@cs.ox.ac.uk;
camiguel@uc.cl

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

Received: *15 January 2020*
Accepted: *09 September 2020*
Published: *21 October 2020*

Citation:
Signorelli CM, Dündar-Coecke S, Wang V and Coecke B (2020) Cognitive Structures of Space-Time. Front. Psychol. 11:527114. doi: 10.3389/fpsyg.2020.527114

1. INTRODUCTION

This article posits a hierarchy in the cognition of spacetime, analogous to a "layer cake" structure, where layers correspond to different aspects of causality. The foundations of the layer-cake structure are derived from physical accounts of causality, supported by a brief mathematical background. The proposed hierarchy acknowledges that neither space nor time can be accessed directly; we can only glean their structures by observing and interacting with objects among events. Therefore, the natural question is how we establish coherent models of spacetime.

Toward an answer, the present paper proposes that cognitive models are hierarchical, where lower layers encode structurally simpler data than higher ones, and the structure of spacetime emerges from mutual constraints between layers.

We take the most primitive layers to be topological, which refers to whether objects and events are "connected." Topology does not distinguish between the types of the lines (e.g., curved or straight); only connectedness—however defined—and its absence, disconnectedness, need be perceived. In the perception of spatial-temporal entities, connectivity and disconnectivity compositionally characterize more complex features such as being "before," "after," "in front," "behind," "having holes," "discreteness," etc.

A more complex, computationally dense and higher up layer might construct metric spaces and Euclidean structure. An example of a constraint between topology and metrics that may arise in some setting is *"objects are connected if and only if they have zero distance from each other."*

Investigating the cognitive structures of space-time governing causal cognition is central to the understanding of a general theory of intelligence in humans and in artificial beings. Nevertheless, in psychology, research lags in providing a concise and systematic review for the correspondences between empirical causal structures and spatial-temporal cognition.

Beyond that, the layer-cake organization of spatial-temporal structures are preserved among other fields, such as physics, mathematics and also computer science, leading to a natural hierarchical organization from topological space (less complex), to metric spaces (more complex). In the following sections, we explore this toy model in the context of physical causal structures (section 2), then provide psychological models (section 3) and continue with a discussion of its implications in a wider context (section 4).

2. LAYERS OF STRUCTURE IN PHYSICS

In Physics, the analysis of the spaces representing potential states of physical systems often takes the form of a layer-cake of increasingly rich structure. The layer-cake is not merely a mathematical decomposition, but is informed by some conceptual underpinning: such as how agents interact with the subject matter, and more specifically, how the subject matter enables/restricts this interaction, or how the subject matter interacts with itself.

A first example is the analysis of relativistic space-time structure as for example in Geroch (2013) and Ehlers et al. (2012). Here, the levels arise from how agents interact with space-time. In Geroch (2013), like in many other such approaches, the first layer is called *causal structure* (**Figure 1A**). It arises from the light-cones that specify which points of space-time (in the future) the agent can affect, and which points of space-time (in the past) the agent can be affected by. Mathematically, these light-cones give rise to a partial order $(P, \leq)$, where for $a, b \in P$ we have $a \leq b$ if space-time point a can affect space-time point b. Often this partial order is taken as a starting point for the development of new physics, for example, when studying quantum causality (Fritz, 2014; Henson et al., 2014), and even when crafting theories of quantum gravity (Bombelli et al., 1987; Sorkin, 2003). A second layer arises from the notion of a clock (**Figure 1A**), which measures the progress of time and hence provides a temporal metric structure atop the partial order of events. Next comes the full space-time metric, followed by dynamical data, among others.

Moving from relativity to quantum theory (QT), following John von Neumann (von Neumann, 1932; Birkhoff and von Neumann, 1936), the first layer is again a partial order, where ordering encodes entailment with respect to agents observing properties of quantum systems, that is, $a \leq b$ if observation of property a guarantees observation of property b. The following layers include conceptually informed universal algebraic equational structure (Piron, 1976). Note that also the entailment relationships can be viewed as a form of informational/epistemic causal structure, as it involves a guaranteed observation given a premise. This branch of quantum theory has mostly vanished from current activity within physics, but has been adopted within psychology in the field of quantum cognition (Busemeyer and Bruza, 2012).

Much more recently, in the category-theoretical analysis of quantum theory (Abramsky and Coecke, 2004; Coecke and Duncan, 2011; Coecke and Kissinger, 2017), rather than the interaction of agents with the subject matter, the lower levels of the layer-cake are informed by how the subject matter interacts with itself. This lowest level is fundamentally *topological*, and more specifically, what topologists call low-dimensional topology (in fact, as low-dimensional as its gets). The structure only expresses what is connected and what is not, without bringing any other geometric notions into play. In this approach, explicit graphical wiring at once formulates and represents connectivity, so it suffices to understand the concept of "wire" to understand this lowest layer of quantum theory (**Figure 1B**). This, in fact, leads to an alternative justification for having this particular layer as the basis: wires are, *a priori*, conceptually primitive for human reasoners (Coecke, 2005, for the indication from the title, namely "Kindergarten quantum mechanics"). An educational experiment is expected to take place during 2020 (see Coecke, 2009), aiming to show that quantum theory presented in topological terms would enable high-school students not only pass a graduate-level quantum theory exam,

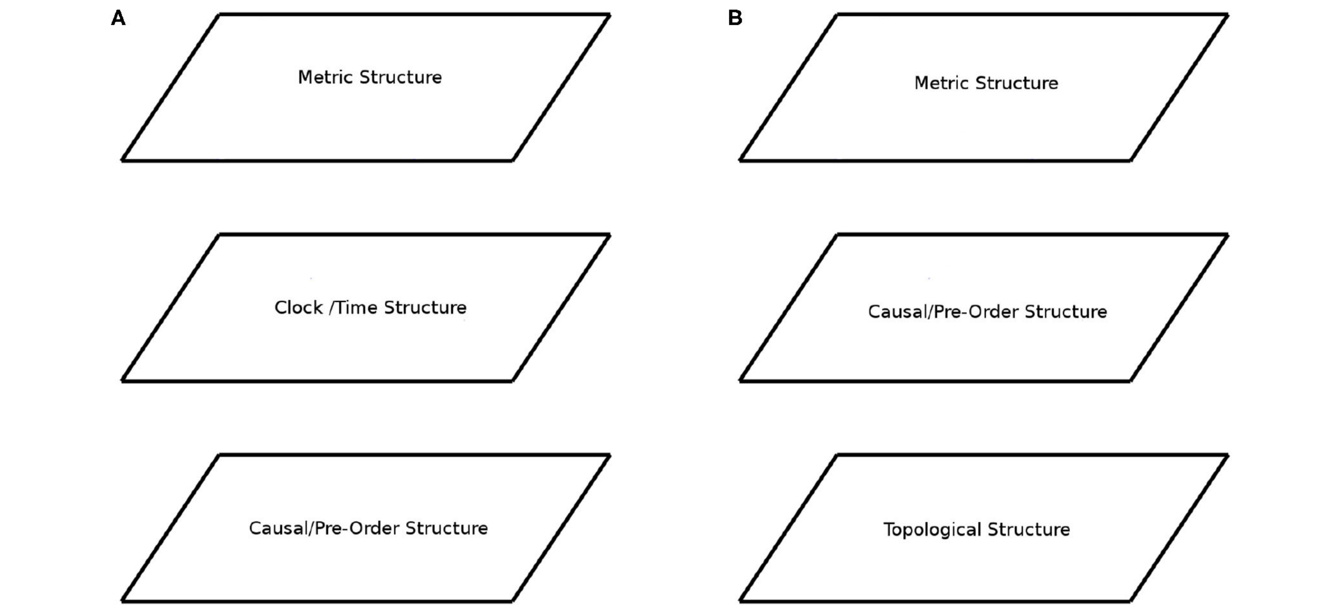


FIGURE 1 | Layer-cake structure. **(A)** The layer structure of Relativistic Space-Time. **(B)** The layer structure of Causal process theories and the hypothesized layer Causal structure for Cognition.

but even outperform university students who are taught the conventional presentation.

Within the topological approach, the notion of causality has been proven to be equivalent to the relativistic notion of causality (Kissinger et al., 2017). Thus causality can be formulated higher up in the layer-cake (Coecke and Kissinger, 2017), synthesized and restrained by more primitive data (**Figure 1B**). In fact, there are multiple presentations on the move from lower topological level to full-blown quantum-theory, cf. Coecke and Kissinger, 2017; Selby et al., 2018, but the topological level is always the beating heart of this approach. As it turns out, natural language is governed by exactly the same topological structures, the reason being that the structure of grammar itself (Lambek, 2008), exactly matches the topological structures of QT (Coecke, 2013, 2017). Furthermore, even more general cognitive models appealing to a wide range of human senses have been shown to be governed by the same structures (Bolt et al., 2018). The starting point here were Gärdenfors (2014)'s conceptual spaces, which aim to closely resemble human senses, and the interaction of these senses is again governed by basic topological structures.

3. LAYERS OF STRUCTURE IN COGNITION

According to previous considerations, cognition may mirror the physical structures of spacetime, or the physical structures suggested by human theories may only reflect a basic cognitive structure of human thinking[1]. Independently of these two options, the layer-cake structures given by physical theories seem to be present in our developmental understanding of spatial and spatial-temporal structure (section 3.2). Therefore in this section, a layer-cake model is discussed as hierarchical levels of cognitive complexity, inheriting, to some extent, all the mathematical properties coming from previous developments in physics DisCoCat/InConcSpec (Coecke et al., 2010; Bolt et al., 2018), without having to develop a new one.

The layer-cake hypothesis addresses a gap in the ongoing neurocognitive debate concerning the—as Bellmund et al. (2018) argue, central—role of spatial-temporal cognition, topology, and metrics in high-level cognition. Direct correlates of euclidean space and time have been identified in neural representation (Moser et al., 2015; Tsao et al., 2018). However, as Buzsáki and Llinás (2017) and Buzsáki and Tingley (2018) observe, the reasoner only receives information concerning distance and duration, reflected in a succession of neuronal events that may not correlate with any space-time representation. This spurs a search for model-building and inferential explanations of how direct neural correlates to space and time arise from sense data, which the layer-cake hypothesis may potentially provide a framework for. Bottini and Doeller (2020) suggest that any such framework goes toward explaining a general propensity of the mind to create low-dimensional internal models. Promisingly, Haun and Tononi (2019) have derived mathematical models demonstrating that brain areas with grid-like connectivity are sufficient to entertain the topological and causal structures necessary for subjective spatial experience. So the layer-cake hypothesis, in concord with all parties of the debate, could serve as a missing link between the mechanical, theoretical, and phenomenal aspects of spatial-temporal cognition.

3.1. Topological Layers of Cognition

The model presented here is a general framework to develop specific implementations according to requirement. The main ingredients are the division/synthesis of causal structure in terms of more primitive structure, and organizing these composite structures into layers corresponding to constraints and affordances of causal relations, and the developmental order.

We propose that the first layer compounds Topological relations, and consequently, that comprehension of causal relations across space and time prioritizes topological structures. It implies that early or primitive forms of causal cognition and specifically spatial cognition would not be highly conceptual, only involving simple notions of proximity, separation, order, enclosure, connectivity, and boundedness. As discussed later, such conceptualization may be through non-symbolic category formation where subjects have restricted access to verbal codes: for example, fundamental ideas about space are developed in infancy by motor and perceptual mechanisms and rely strongly upon sensory/perceptual data. Diagrammatically, two objects *A* and *B*, are topologically related if there is an event that connects them, which is defined by the relation *R(A,B)*. These connections are usually described by wires and objects by nodes. Under this notation, wires are relational events and circles are static objects (**Figure 2**).

The relation events *R(A,B)* and *S(C,D)* connecting the objects of cognition described by *A*, *B*, *C*, and *D* correspond to fundamental and basic notions, that eventually lead to the understanding of spatial relations. Later, other types of relation emerge, such as the effects between objects, which correspond to object interactions across primitive notions of time. These interactions define processes notated by boxes, such as *f*. More specifically, such interactions may correspond to a causal processes according to a partial order relation (**Figure 2A**). In other words, the object *A* and *B* become causally related systems under the partial order, written $A \leq B$, meaning in the abstract that information flows unidirectionally from *A* to *B*, thus defining a second layer of structure upon systems. Notably, causal relations defined in this way among objects are not necessarily unique, as exemplified by the case of *C* and *D*. Following the notation from previous works (Coecke and Kissinger, 2017), now wires become objects/systems and boxes the causal processes among them (**Figure 2A**).

Empirically, we abduct events from *observations* of relational spatial properties. In contrast, *processes* may encompass unobservable intervening dynamical factors (e.g., forces), which need to be constructed or reconstructed in further levels of complexity: processes correspond to abstract components of mechanisms. Therefore, the second layer would correspond to the representational/relations space associated with causal interactions, governed by the partial order relations mentioned above. We hypothesize that the gradual emergence of concepts, syntax, grammar would be associated with such higher layers, as

[1] That issue together with the possible neural realization will be discussed elsewhere.

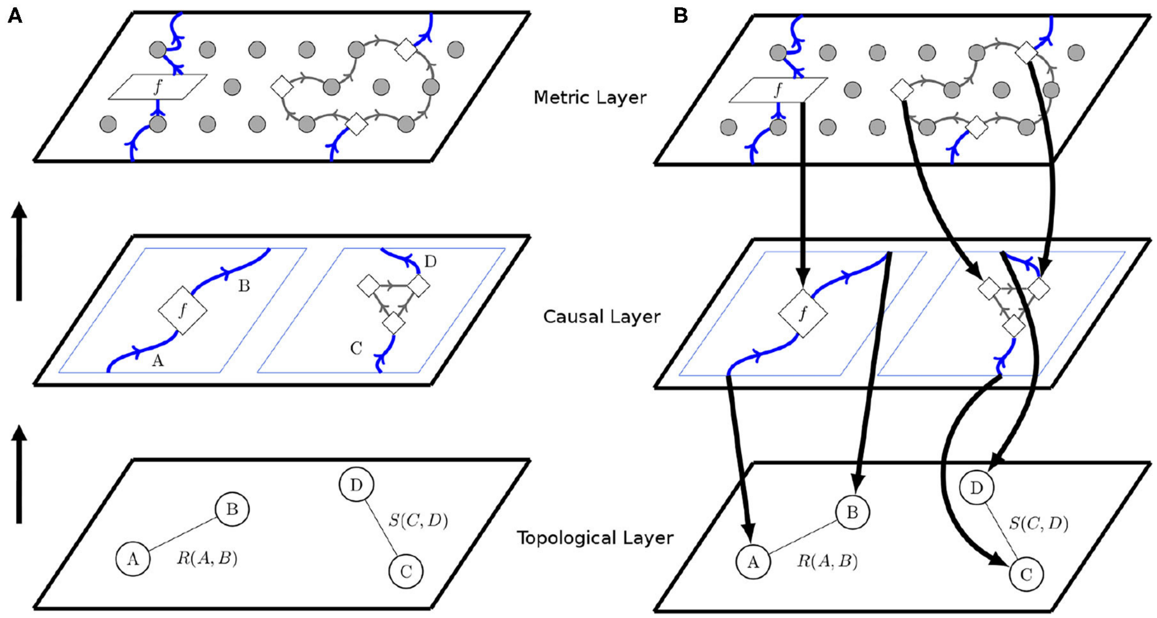


FIGURE 2 | Layer structure for cognition. **(A)** Topological Structure is relational and increase in complexity as we rise through the layers. In the topological layer, circles represent objects and wires topological relations between objects. Then, causal structure, in the form of process theories is build on top of topological relations from the lower layer. In this second pre-order or basic causal layer, objects become represented by wires and causal relations by processes. Finally, the metric layer adds metric structure, which elaborates causal structure in a suitable, spatial and temporal setting. This condition is given by the black dots in the figure. **(B)** This layer division generates a hierarchical structure, where higher layers are structurally constrained by the data of lower layers as well as they can influence part of the lower configuration.

these permit representation and reasoning with counterfactual and imaginary phenomena not immediately constrained by past experience and direct perception.

A consequence of this division is that constraint-satisfying structure on any layer, in turn, places constraints on how further layers are defined. Viewing foundational layers as abstract schema or cognitive resources (and their neural realizations) shapes the modes of access to that structure, constraining how relations take place in that schema. For instance, when we take processes in spacetime to be mutually exclusive, we can begin to fill in complex narratives. If we know that a battle and a wedding took place in the same valley, mutual exclusivity of processes and linear temporal ordering allow us to raise a fruitfully constrained set of alternative models: either the battle came before the wedding, or *vice versa*.

Hence, any layer may be viewed as an abstract space upon a lower layer, the higher further specifying instances of structure compatible with those of the lower (**Figure 2B**). In **Figure 2A**, the higher layer carries the particular refinement of metric structure. The precise nature of cognitive metric structures is a question for future research, and not our chief concern here. No matter the metric, according to the layer-cake model, representation and reasoning in metric spaces is more computationally intensive than in topological spaces, because higher layers carry a greater informational capacity than lower ones, and carry more constraints and affordances for the reasoner to navigate.

These emergent hierarchies are subjective to the reasoner, and not an objective feature of reality: hence, we can speak distinctly of perceived vs. objective causality. In other words, while the seemly real characteristics of spacetime affect how we conceptualize spacetime, our conceptualization in turn dynamically constrains and directs further conceptualization.

Finally, a word of caution when interpreting the topological hypothesis as stated above is that different conceptions of causality and topology exist, as these are not uniquely defined concepts across disciplines, and not even in pure mathematics, where a field like topology has several very different branches of study that are qualitatively different. For example, taking path-connectedness as the primitive—where one identifies possible paths that one can take between points in space—will cause one to identify all points on the surface of a table as "*essentially the same*," whereas homology theory—where one identifies the characterizing holes of a structure—will cause one to treat drinking mugs and donuts as "*essentially the same*." The layer-cake model accommodates any and all particular formulations of topology, as it is synthetic: the fundamental ingredient of defining higher structures atop lower ones remains in play.

3.2. Supporting Evidence for the Layer-Cake Structure

Developmental studies are in accord with the layer-cake hypothesis. Evidence supports the notion that topological properties, representing the earliest/primitive forms of distance-duration relations, are available initially through a nonverbal category formation, even where young children have restricted access to verbal codes (Dündar-Coecke et al., 2020). Using linguistic and non-linguistic tasks (Piaget, 1959) (see also Piaget and Inhelder, 1971) pioneered the argument that infants' perceptual space is qualitatively different from that of adults. At the beginning, fundamental spatial concepts are not Euclidean, but topological, which involves some concepts such as proximity, separation, order, enclosure, long before it becomes metric. This suggests that the infant's space must be quite fluid, not objective, nor occupied by rigid shapes or sizes.

Studies of adult cognition also acknowledge this fluency in cognitive structures. In Biederman and Cooper (1991) and Biederman and Cooper (1992) study, although participants were presented with contour-deleted pictures, they completed perceptual stimuli in the absence of size, location, or orientation information, highlighting humans' ability to recognize objects independent of Euclidian spatial features in a more abstract fashion. While these results suggest a potential primacy of topology over more complex data, research establishing cognitive mechanisms involved in conceptualization of topological and metric properties also provides consistent evidence that people cannot act within, or orient themselves to their environments unless provided spatial and temporal information constituting their physical reality (Han et al., 1999; Müsseler, 1999; Chen, 2005).

Topology's fundamental role in understanding space is supported by theoretical grounds in neuroscience: Marr (1982) posits a sophisticated *motion correspondence process* in the perception of an entity through time, simple topological transformations also enable observation of apparent motion (Chen, 1982, 2005; Ogmen and Herzog, 2010). Rock and Palmer (1990) stress the law of "connectedness" in early perceptual analysis, and the topological perception hypothesis suggests that shape-changing transformations experienced in the phenomenal world rely on topological transformations, for example, projected in retina with the aid of three kinds of topological properties: connectivity, the number of holes, and the inside/outside relationship.

Another strand of work emphasizes the role of selective attention as a strategy to bias continually registerable spatial-temporal attributes, and hence increase control in processing capacities through top-down neural connections (Kastner and Ungerleider, 2000). In fact, neuroimaging studies have shown that a number of mechanisms can contribute to attentional orientation to moving targets (Doherty et al., 2005; Shimi et al., 2014), with a prevailing view that perceptual organization (topological) likely to occur before feature analysis (metrics). Chen (1982) reports a series of experimental findings showing the precedence of topological feature detection in the visual system, further supporting the view that topological features form conceptual foundations. Pomerantz's configural superiority effect supports this hypothesis (Pomerantz, 1981; see also Todd, 1998), by adding that features can be observed even in response to stimuli that are not fully configural, as configural information is already present at early stages of visual hierarchy (see also Fox et al., 2017, for neural evidence).

Limited knowledge in furthering these discussions urges us to swing the pendulum back to the infant studies, which are highly informative regarding the detection of primitive forms of spatial-temporal properties. Infants appear to show sensitivity to moving objects along "continuous" paths, and also pay attention to interactions only if they are causally in contact (see Leslie, 1984; Leslie and Keeble, 1987; Spelke et al., 1992, 1995a; Spelke, 1994, see also Darcheville et al., 1993, for how infants learn about space as a function of the temporal intervals). However, they seem to find it difficult to relate objects based on non-causal qualities, such as colors, forms, edges, or surfaces (Kellman and Spelke, 1983). Instead, they show a tendency to rely on simple forms of spatial-temporal information to distinguish different types of objects and events (see Slater et al., 1994; Spelke et al., 1995b; Needham et al., 1997; Wilcox and Baillargeon, 1998; see also Kaufman et al., 2003, for evidence how spatial-temporal stimuli are processed by different visual streams). These studies propose consistent evidence for the early sensitivity to topological spatial-temporal features such as continuity and connectivity in causal contexts. Although maturation in use of these representations are accompanied by conceptual development, humans are multimodal reasoners; most implicit spatial-temporal qualities are more akin to sensations and do not necessarily have to be available to communication (Tolmie and Dündar-Coecke, 2020). This may explain the consistency between adult and infant data.

The early fundamental ideas about space-time develop largely by embodied motor and sensory activities. Young children experience the most primitive spatial-temporal properties via observing, touching, and moving their/others' bodies. The development of symbolic cognitive resources, such as memory and language, enables spatial-temporal properties to become more representational, allowing children to mentally evoke objects in their physical absence. Understanding of or paying attention to metrics and Euclidean structures emerge as a function of the development of these internal and external resources and models. For instance, a child learns how to stack the smaller object into the big ones, or improve projective and perspective taking skills gradually. *Contextual* consistency of spatial models appears to develop later than spatial models of individual closed objects. For example, at early stages, children are likely draw a human being bigger than e.g., a house in size, while the orientation of both human and house may/not respect gravity, and relative placement of appendages and windows all correct for both human and house. The primary context in which size consistency is obtainable is the embodied motor-sensory paradigm: at the same physical distance from a human and a house, the human image may have a smaller angle of subtension in the infant's field of vision.

Therefore, developmental literature underlines the myriad ways in which spatial-temporal properties are experienced and employed in service of causal cognition, in accord with the layer-cake hypothesis where causal relations are predicated upon spatial-temporal foundation layers. The most studied spatial-temporal attributes in causal cognition literature are properties high in the layer-cake: distance, duration, velocity, and spatial-temporal incongruences (Bullock and Gelman, 1979; Siegler and Richards, 1979; Wilkening, 1981; Bullock et al., 1982; Wilkening and Cacchione, 2011). These studies sample either children or adults, and a comparison between these and early infancy studies implies that the more children/humans are able to utilize spatial-temporal properties in Euclidian fashion, the better they can acknowledge causal relations. Although the grasp of causal relations requires the organization of connections across space and time in topological sense and this is critical for visual function at any age, the genuine understanding of cause-effect relations matures when we define the richer causal geography of spacetime.

4. CONCLUSIONS AND FUTURE RESEARCH

The layer-cake hypothesis provides a meta-model of spacetime cognition. The main argument of this conceptual model is that spatial and temporal qualities increase in their complexity across mutually constraining layers of description, ranging from the topological to metric, temporal, and causal, for models of physical or virtual/abstract spaces. It is the layer-cake taken as a whole that can be considered the full model. The hierarchical organization of layers is a novel form to study this complexity of the spatial-temporal relations in both physics and psychology, providing rich enough model to capture not only the interaction of multiple dimensions of abstractions, but the internal dynamics of constructing cognitive models from empirical data, fed by the reciprocal interactions between perception, action, and reasoning about space, time and causality.

The layer-cake hypothesis is adaptable but crucially for science, defeasible, as it must always be instantiated to provide concrete models. These instantiations compatibly formalize a broad range of current approaches to cognition of causality across space and time. Previously, Newcombe and Shipley (2015) and Uttal et al. (2013) studies underlined how the intrinsic/extrinsic and static/dynamic relations between entities inform us about the characteristics of spatial elements, which may be modeled as graphical calculi on suitably encoded layers of a layer-cake. Developmental origins of thinking about past, current, future situations (Friedman, 2003; McCormack and Hoerl, 2005), either in segmented, speeded, or imagined protocols (Dündar-Coecke et al., 2020) may be formalized in the physicist's language of logics upon partial orders on events, again amenable to graphical and layer-cake methods of representation and reasoning. Layer-cake models are well-suited to novel developmental studies in calibration and approximation of spatial-temporal attributes on virtual displays (Dündar-Coecke, 2019), where the spatial environment is distanced from the young reasoner by a layer of abstract representation, as layer-cakes have tunable levels of abstraction built-in.

On the theoretical side, our perspective aims to generate a new interdisciplinary semantics for spatio-temporal cognitio interwoven with theoretical physics. In conjunction with experimental phases, if the layer-cake structure deduced from theoretical physics is shared or preserved in the structure of spatio-temporal cognition, we can shed light on those structures using recent mathematical tools that deal with physical space-time and causality. Throughout, we expect to use axiomatic process-theoretical tools which are currently applied for causal relationships in physics (Coecke and Kissinger, 2017; Kissinger and Uijlen, 2017). This approach will allow us to describe the nature of spatio-temporal experience in the form of interacting processes, following similar strategies already implemented for language and cognition (Coecke et al., 2010, 2018; Coecke, 2013; Bolt et al., 2018).

On the experimental side, one can ask about the neural and behavioral implications of our axiomatic models. For example, if we establish the presence of distinct but cohesive competencies for different aspects of spatial cognition and experience, a subsequent question is to ask where does the layer-cake find expression? The question of whether this paradigm finds implementational reality inside brains (as suggested by Signorelli, 2018; Signorelli and Meling, 2020) and the discussion of the feasibility of layer-cake models in terms of neural structure will form part of further extensions to this program. More broadly, we may unlock spaces of questions for developmental and evolutionary biology, to further our understanding of how agents arise in space-time and vice versa.

AUTHOR CONTRIBUTIONS

CS conceptualization and visualization the model, writing the original manuscript, and editing subsequent versions. SD-C conceptualization the model and wrote and edited the manuscript. VW conceptualization the model and edited subsequent versions. BC conceptualization the model and writing manuscript. All authors contribute to the original hypothesis and discussions.

FUNDING

This publication was made possible through the support of the ID# 61466 grant from the John Templeton Foundation, as part of the The Quantum Information Structure of Spacetime (QISS) Project (qiss.fr). The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. This work was partially supported by the award ES/P000592/1 from the UK ESRC; and Comisión Nacional de Investigación Ciencia y Tecnología (CONICYT) through Programa Formacion de Capital Avanzado (PFCHA), Doctoral scholarship Becas Chile: CONICYT PFCHA/DOCTORADO BECAS CHILE/2016 - 72170507.

ACKNOWLEDGMENTS

The authors appreciate the encouraging comments and suggestions from the reviewers.

REFERENCES

Abramsky, S., and Coecke, B. (2004). "A categorical semantics of quantum protocols," in *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science (LICS)* (Oxford), 415–425. doi: 10.1109/LICS.2004.1319636

Bellmund, J. L. S., Gardenfors, P., Moser, E. I., and Doeller, C. F. (2018). Navigating cognition: spatial codes for human thinking. *Science* 362:6415. doi: 10.1126/science.aat6766

Biederman, I., and Cooper, E. E. (1991). Priming contour-deleted images: evidence for intermediate representations in visual object recognition. *Cogn. Psychol*. 23, 393–419. doi: 10.1016/0010-0285(91)90014-F

Biederman, I., and Cooper, E. E. (1992). Size invariance in visual object priming. *J. Exp. Psychol. Hum. Percept. Perform*. 18, 121–133. doi: 10.1037/0096-1523.18.1.121

Birkhoff, G., and von Neumann, J. (1936). The logic of quantum mechanics. *Ann. Math*. 37, 823–843. doi: 10.2307/1968621

Bolt, J., Coecke, B., Genovese, F., Lewis, M., Marsden, D., and Piedeleu, R. (2018). "Interacting conceptual spaces I: grammatical interaction of concepts," in *Concepts and their Applications, Synthese Library, Studies in Epistemology, Logic, Methodology, and Philosophy of Science*, eds M. Kaipainen, A. Hautamaki, P. Gardenfors, and F. Zenker (Springer). doi: 10.1007/978-3-030-12800-5_9

Bombelli, L., Lee, J., Meyer, D., and Sorkin, R. (1987). Space-time as a causal set. *Phys. Rev. Lett*. 59:521. doi: 10.1103/PhysRevLett.59.521

Bottini, R., and Doeller, C. F. (2020). Knowledge across reference frames: cognitive maps and image spaces. *Trends Cogn. Sci*. 24, 606–619. doi: 10.1016/j.tics.2020.05.008

Bullock, M., and Gelman, R. (1979). Preschool children's assumptions about cause and effect: temporal ordering. *Child Dev*. 50, 89–96. doi: 10.2307/1129045

Bullock, M., Gelman, R., and Baillargeon, R. (1982). "The development of causal reasoning," in *The Developmental Psychology of Time*, ed W. J. Friedman (New York, NY: Academic Press), 209–254.

Busemeyer, J. R., and Bruza, P. D. (2012). *Quantum Models of Cognition and Decision*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511997716

Buzsáki, G. and Llinás, R. (2017). Space and time in the brain. *Science* 358, 482–485. doi: 10.1126/science.aan8869

Buzsáki, G., and Tingley, D. (2018). Space and time: the hippocampus as a sequence generator. *Trends Cogn. Sci*. 22, 853–869. doi: 10.1016/j.tics.2018.07.006

Chen, L. (1982). Topological structure in visual perception. *Science* 218, 699–700. doi: 10.1126/science.7134969

Chen, L. (2005). The topological approach to perceptual organization. *Vis. Cogn*. 12:553–637. doi: 10.1080/13506280444000256

Coecke, B. (2005). "Kindergarten quantum mechanics," in *Quantum Theory: Reconsiderations of the Foundations III*, ed A. Khrennikov (Växjö: AIP Press), 81–98. doi: 10.1063/1.2158713

Coecke, B. (2009). Quantum picturalism. *Contemp. Phys*. 51, 59–83. doi: 10.1080/00107510903257624

Coecke, B. (2013). "An alternative Gospel of structure: order, composition, processes," in *Quantum Physics and Linguistics. A Compositional, Diagrammatic Discourse*, eds C. Heunen, M. Sadrzadeh, and E. Grefenstette (Oxford: Oxford University Press), 1–22.

Coecke, B. (2017). "From quantum foundations via natural language meaning to a theory of everything," in *The Incomputable: Journeys Beyond the Turing Barrier, Theory and Applications of Computability*, eds S. B. Cooper and M. I. Soskova (Springer International Publishing), 63–80. doi: 10.1007/978-3-319-43669-2_4

Coecke, B., and Duncan, R. (2011). Interacting quantum observables: categorical algebra and diagrammatics. *N. J. Phys*. 13:043016. doi: 10.1088/1367-2630/13/4/043016

Coecke, B., Genovese, F., Lewis, M., Marsden, D., and Toumi, A. (2018). Generalized relations in linguistics & cognition. *Theoret. Comput. Sci*. 752, 104–115. doi: 10.1016/j.tcs.2018.03.008

Coecke, B., and Kissinger, A. (2017). *Picturing Quantum Processes. A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge: Cambridge University Press. doi: 10.1017/9781316219317

Coecke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *arXiv [Preprint]*. arXiv:1003.4394.

Darcheville, J., Riviare, V., and Wearden, J. (1993). Fixed-interval performance and self-control in infants. *J. Exp. Anal. Behav*. 60, 239–254. doi: 10.1901/jeab.1993.60-239

Doherty, J., Rao, A., Mesulam, M., and Nobre, A. (2005). Synergistic effect of combined temporal and spatial expectations on visual attention. *J. Neurosci*. 25, 8259–8266. doi: 10.1523/JNEUROSCI.1821-05.2005

Dündar-Coecke, S. (2019). "Do children represent virtual spatial-temporal qualities different than adults?" in *Cognition and Exploratory Learning in Digital Age* (Cagliari). doi: 10.33965/celda2019_201911L046

Dündar-Coecke, S., Tolmie, A., and Schlottmann, A. (2020). The role of spatial and spatial-temporal analysis in children's causal cognition of continuous processes. *PLoS ONE* 15:e0235884. doi: 10.1371/journal.pone.0235884

Ehlers, J., Pirani, F. A., and Schild, A. (2012). Republication of: The geometry of free fall and light propagation. *Gen. Relativ. Gravit*. 44, 1587–1609. doi: 10.1007/s10714-012-1353-4

Fox, O., Harel, A., and Bennett, K. (2017). How configural is the configural superiority effect? A neuroimaging investigation of emergent features in visual cortex. *Front. Psychol*. 8:32. doi: 10.3389/fpsyg.2017.00032

Friedman, W. J. (2003). *The Development of a Differentiated Sense of the Past and Future. Volume 31 of Advances in Child Development and Behaviour*. San Diego, CA: Academic Press.

Fritz, T. (2014). Beyond Bell's theorem II: scenarios with arbitrary causal structure. *arXiv [Preprint]*. arXiv:1404.4812. doi: 10.1007/s00220-015-2495-5

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9629.001.0001

Geroch, R. (2013). *General Relativity: 1972 Lecture Notes*, Vol. 1. Montreal, QC: Minkowski Institute Press.

Han, S., Humphreys, G., and Chen, L. (1999). Parallel and competitive processes in hierarchical analysis: perceptual grouping and encoding of closure. *J. Exp. Psychol. Hum. Percept. Perform*. 25, 1411–1132. doi: 10.1037/0096-1523.25.5.1411

Haun, A., and Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy* 21:12. doi: 10.3390/e21121160

Henson, J., Lal, R., and Pusey, M. F. (2014). Theory-independent limits on correlations from generalised Bayesian networks. *arXiv preprint arXiv:1405.2572*. doi: 10.1088/1367-2630/16/11/113043

Kastner, S., and Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci*. 23, 315–341. doi: 10.1146/annurev.neuro.23.1.315

Kaufman, J., Mareschal, D., and Johnson, M. (2003). Graspability and object processing in infants. *Infant Behav. Dev*. 26, 516–528. doi: 10.1016/j.infbeh.2002.10.001

Kellman, P., and Spelke, E. (1983). Perception of partly occluded objects in infancy. *Cogn. Psychol*. 15, 483–524. doi: 10.1016/0010-0285(83)90017-8

Kissinger, A., Hoban, M., and Coecke, B. (2017). Equivalence of relativistic causal structure and process terminally. *arXiv [Preprint]*. arXiv:1708.04118.

Kissinger, A., and Uijlen, S. (2017). "A categorical semantics for causal structure," in *32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)* (Reykjavík), 1–12. doi: 10.1109/LICS.2017.8005095

Lambek, J. (2008). *From Word to Sentence*. Milan: Polimetrica.

Leslie, A. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception* 13, 287–305. doi: 10.1068/p130287

Leslie, A., and Keeble, S. (1987). Do six-month-olds infants perceive causality? *Cogniti* Polimetrica *on* 25, 265–288. doi: 10.1016/S0010-0277(87)80006-9

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

McCormack, T., and Hoerl, C. (2005). Children's reasoning about the causal significance of the temporal order of events. *Dev. Psychol*. 41, 54–63. doi: 10.1037/0012-1649.41.1.54

Moser, M.-B., Rowland, D. C., and Moser, E. I. (2015). Place cells, grid cells, and memory. *Cold Spring Harbor Perspect. Biol*. 7:a021808. doi: 10.1101/cshperspect.a021808

Müsseler, J. (1999). "Perceiving and measuring of spatiotemporal events," in *Modeling Consciousness across the Disciplines*, ed S. J. Jordan (Lanham, MD: University Press of America), 95–112.

Needham, A., Baillargeon, R., and Kaufman, L. (1997). *Object Segregation in Infancy. Volume 11 of Advances in Infancy Research*. Greenwich, CT: Ablex.

Newcombe, N., and Shipley, T. (2015). "Thinking about spatial thinking: new typology, new assessments," in *Studying Visual and Spatial Reasoning for Design Creativity*, ed J. S. Gero (Dordrecht: Springer), 1–18. doi: 10.1007/978-94-017-9297-4_10

Ogmen, H., and Herzog, M. H. (2010). The geometry of visual perception: retinotopic and non-retinotopic representations in the human visual system. *Proc. IEEE* 98, 479–492. doi: 10.1109/JPROC.2009.2039028

Piaget, J. (1959). *The Language and Thought of the Child, 3rd Edn*. New York, NY: Routledge and Kegan Paul.

Piaget, J., and Inhelder, B. (1971). *The Child's Conception of Space*. London: Routledge and Kegan Paul.

Piron, C. (1976). "On the foundations of quantum physics," in *Quantum Mechanics, Determinism, Causality, and Particles. Mathematical Physics and Applied Mathematics*, Vol. 1, eds M. Flato, Z. Maric, A. Milojevic, D. Sternheimer and J. P. Vigier (Dordrecht: Springer). doi: 10.1007/978-94-010-1440-3_7

Pomerantz, J. R. (1981). "Perceptual organization in information processing," in *Perceptual Organization*, eds M. Kubovy and J. R. Pomerantz (Hillsdale, NJ: Lawrence Erlbaum), 141–180. doi: 10.4324/9781315512372-6

Rock, I., and Palmer, S. (1990). The legacy of gestalt psychology. *Sci. Am*. 263, 84–90. doi: 10.1038/scientificamerican1290-84

Selby, J. H., Scandolo, C. M., and Coecke, B. (2018). Reconstructing quantum theory from diagrammatic postulates. *arXiv [Preprint]*. arXiv:1802.00367.

Shimi, A., Kuo, B., Astle, D., Nobre, A., and Scerif, G. (2014). Age group and individual differences in attentional orienting dissociate neural mechanisms of encoding and maintenance in visual STM. *J. Cogn. Neurosci*. 26, 864–877. doi: 10.1162/jocn_a_00526

Siegler, R. S., and Richards, D. D. (1979). Development of time, speed, and distance concepts. *Dev. Psychol*. 15, 288–298. doi: 10.1037/0012-1649.15.3.288

Signorelli, C., and Meling, D. (2020). Towards new concepts for a biological neuroscience of consciousness. *PsyArXiv [Preprint]*. doi: 10.31234/osf.io/pcmj9

Signorelli, C. M. (2018). Can computers become conscious and overcome humans? *Front. Robot. AI* 5:121. doi: 10.3389/frobt.2018.00121

Slater, A., Johnson, S. P., Kellman, P. J., and Spelke, E. S. (1994). The role of three-dimensional cues in infants' perception of partly occluded objects. *Early Dev. Parent*. 3, 187–191. doi: 10.1002/edp.2430030308

Sorkin, R. D. (2003). *Causal Sets: Discrete Gravity. Notes for the Valdivia Summer School.* Valdivia, Report number SU-GP-2003/1-2.

Spelke, E., Phillips, A., and Woodward, A. (1995a). *Infants' Knowledge of Object Motion and Human Action. Causal Cognition: A Multidisciplinary Debate*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198524021.003.0003

Spelke, E. S. (1994). Initial knowledge: six suggestions. *Cognition* 50, 431–445. doi: 10.1016/0010-0277(94)90039-6

Spelke, E. S., Breinlinger, K., Macomber, J., and Jacobson, K. (1992). Origins of knowledge. *Psychol. Rev*. 99, 605–632. doi: 10.1037/0033-295X.99.4.605

Spelke, E. S., Kestenbaum, R., Simons, D. J., and Wein, D. (1995b). Spatiotemporal continuity, smoothness of motion and object identity in infancy. *Brit. J. Dev. Psychol*.13, 113–143. doi: 10.1111/j.2044-835X.1995.tb00669.x

Todd, J. T. (1998). "Theoretical and biological limitations on the visual perception of three-dimensional structure from motion," in *High-Level Motion Processing: Computational, Neurophysiological and Psychophysical Perspectives* (Cambridge, MA: MIT Press), 359–380.

Tolmie, A., and Dündar-Coecke, S. (2020). "Lifespan conceptual development in science: brain and behaviour," in *Educational Neuroscience: Development Across the Life Span*, eds M. S. C. Thomas, D. Mareschal, and I. Dumontheil (New York, NY: Routledge), 193–220.

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M.-B., et al. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature* 561, 57–62. doi: 10.1038/s41586-018-0459-6

Uttal, D., Meadow, N., Tipton, E., Hand, L., Alden, A., Warren, C., et al. (2013). The malleability of spatial skills: a meta-analysis of training studies. *Psychol. Bull*. 139, 352–402. doi: 10.1037/a0028446

von Neumann, J. (1932). *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer-Verlag.

Wilcox, T., and Baillargeon, R. (1998). Object individuation in infancy: the use of featural information in reasoning about occlusion events. *Cogn. Psychol*. 37, 97–155. doi: 10.1006/cogp.1998.0690

Wilkening, F. (1981). Integrating velocity, time and distance information: a developmental study. *Cogn. Psychol*. 13, 231–247. doi: 10.1016/0010-0285(81)90009-8

Wilkening, F., and Cacchione, T. (2011). "Children's intuitive physics," in *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, ed U. Goswami (Oxford: Wiley-Blackwell), 473–496. doi: 10.1002/9781444325485.ch18

Conflict of Interest: 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 03 November 2020
doi: 10.3389/fpsyg.2020.502751

Explaining Away, Augmentation, and the Assumption of Independence

Nicole Cruz[1]*, *Ulrike Hahn*[1], *Norman Fenton*[2] *and David Lagnado*[3]

[1] Department of Psychological Sciences, Birkbeck, University of London, London, United Kingdom, [2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom, [3] Department of Experimental Psychology, University College London, London, United Kingdom

In reasoning about situations in which several causes lead to a common effect, a much studied and yet still not well-understood inference is that of *explaining away.* Assuming that the causes contribute independently to the effect, if we learn that the effect is present, then this increases the probability that one or more of the causes are present. But if we then learn that a particular cause is present, this cause "explains" the presence of the effect, and the probabilities of the other causes decrease again. People tend to show this explaining away effect in their probability judgments, but to a lesser extent than predicted by the causal structure of the situation. We investigated further the conditions under which explaining away is observed. Participants estimated the probability of a cause, given the presence or the absence of another cause, for situations in which the effect was either present or absent, and the evidence about the effect was either certain or uncertain. Responses were compared to predictions obtained using Bayesian network modeling as well as a sensitivity analysis of the size of normative changes in probability under different information conditions. One of the conditions investigated: when there is certainty that the effect is absent, is special because under the assumption of causal independence, the probabilities of the causes remain invariant, that is, there is no normative explaining away or augmentation. This condition is therefore especially diagnostic of people's reasoning about common-effect structures. The findings suggest that, alongside earlier explanations brought forward in the literature, explaining away may occur less often when the causes are assumed to interact in their contribution to the effect, and when the normative size of the probability change is not large enough to be subjectively meaningful. Further, people struggled when given evidence against negative evidence, resembling a double negation effect.

Keywords: intercausal reasoning, explaining away, noisy-or, uncertain evidence, negative evidence

OPEN ACCESS

Edited by:
York Hagmayer,
University of Göttingen, Germany

Reviewed by:
Annelie Rothe-Wulf,
University of Freiburg, Germany
Andrew Tolmie,
University College London,
United Kingdom

****Correspondence:***
Nicole Cruz
n.cruz@unsw.edu.au

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

Received: *04 October 2019*
Accepted: *17 September 2020*
Published: *03 November 2020*

Citation:
Cruz N, Hahn U, Fenton N and Lagnado D (2020) Explaining Away, Augmentation, and the Assumption of Independence.
Front. Psychol. 11:502751.
doi: 10.3389/fpsyg.2020.502751

INTRODUCTION

Imagine you are on a tropical island in which there are three types of mosquito (Reb, Mar, and Murb) that carry a disease, called Ling fever. For each mosquito type, there is a risk of being bitten by an infected mosquito, and a risk of contracting the disease when bitten. One day during a routine health check, it turns out that you have Ling fever, prompting you to increase your degree of belief that you were bitten by an infected mosquito. Further tests show that you were bitten by an infected mosquito of the Reb type. How does this additional information affect your degree of belief that you were bitten by an infected mosquito of the Mar type? In this situation, the presence of a bite from

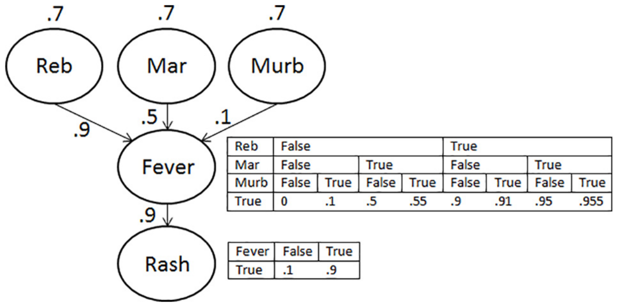
Reb "explains away" the finding of Ling fever, suggesting one can reduce one's degree of belief in a bite from Mar (Rehder and Waldmann, 2017).

Now imagine the further test showed instead that you were *not* bitten by an infected mosquito of the Reb type. How does this additional piece of information affect your degree of belief that you were bitten by an infected mosquito of the Mar type? In the absence of a bite from Reb, the finding of Ling fever is still in need of an explanation, suggesting one can "augment" one's degree of belief in a bite from Mar.

The above reasoning is called intercausal because it involves inferring the likelihood that one cause is present or absent, based on knowledge about one or more further causes. People have been found to show explaining away and augmenting in intercausal reasoning tasks, though not always reliably. In particular the size of the effects has sometimes been smaller than predicted (Morris and Larrick, 1995; Ali et al., 2011; Rottman and Hastie, 2014; Liefgreen et al., 2018; Tešić et al., 2020). The present paper aims to investigate further the conditions under which these inferences are drawn. It explores to what extent people change their intuitions about augmenting and explaining away as a function of (a) whether the evidence about the effect is positive or negative, and (b) whether this evidence is certain or uncertain. But before going into more detail about these two factors, let us turn briefly to the general framework within which changes in people's degrees of belief like those of explaining away and augmentation can be represented.

Changes in degrees of belief over time as new information about a situation becomes available can be modeled in a Bayesian network (BN) (Pearl, 1988, 2000). In a BN, the relevant events are represented as variables and arrows represent (non)independence relations connecting the variables, forming a directed acyclic graph (DAG). Associated with each variable is a conditional probability table (CPT), which specifies the probability of each value that the variable can take, as a function of each of the possible values of the variables on which it directly depends (i.e., is linked to by arrows). In this way, BNs allow the graphical representation and variation of complex probabilistic relations between events, making transparent which variables are positively or negatively related to one another, and which are independent, and supporting the computation of dynamic changes to beliefs as evidence comes in. This probabilistic, Bayesian approach to causal reasoning provides an alternative to earlier approaches based on classical logic (Fernbach and Erb, 2013; Oaksford and Chater, 2017; Over, 2017), possible worlds semantics (Lewis, 1973; Stalnaker, 1981; Briggs, 2012), and theories of associative learning (Waldmann and Holyoak, 1992; Sloman and Lagnado, 2005, 2015; Rehder, 2014).

A BN for our mosquito example has three (marginally independent) causes (a bite of an infected mosquito of type Reb, Mar, or Murb), and one common effect (Ling fever). Such a structure is shown in **Figure 1**. The CPT for the effect would then contain the probability of Ling fever for each combination of the truth or falsity of each of the three causes, yielding eight distinct entries like those shown. People may not have clear intuitions about the probability of each of the eight entries, but fewer parameters need to be specified if one can draw on a more general function specifying how the impact of the causes combines to bring about (or prevent) the effect (c.f. Fenton et al., 2007).



Reb	False				True			
Mar	False		True		False		True	
Murb	False	True	False	True	False	True	False	True
True	0	.1	.5	.55	.9	.91	.95	.955

Fever	False	True
True	.1	.9

FIGURE 1 | Causal structure of the scenario. The upper conditional probability table (CPT) displays the probability of fever given the presence or absence of each of the three causes. The lower CPT shows the probability of a rash given the presence or absence of fever. These CPTs follow from the priors and causal power values shown in the graph, together with a leakage parameter of 0 for Fever, and of 0.1 for Rash.

A typical function for common-effect structures like that of the mosquito example is the *noisy-or*. The noisy-or specifies the probability of the effect given a disjunction of independent causes. It is a generalization of the Boolean OR to reasoning from uncertain premises. The basic idea is that the probability of a disjunction is equal to 1 minus the probability of the negation of the disjunction, so that P(effect| *A or B*) = 1 – P(effect| *not-A & not-B*). Formally, let $x_i = x_1,..,x_n$ be n variables representing the causes of an effect y. Let v_i be a weight factor for each cause, specifying the conditional probability of the effect given cause i in the absence of the other causes (i.e., the causal power of cause *i*, Cheng, 1997). Finally, let λ be a *leakage parameter* specifying the probability that the effect occurs when all the causes included in the model are absent. The leakage parameter is like a residual category covering the impact of any causes that have not been explicitly specified. Then the probability of the effect is given by:

$$\left(y = 1|x_1, \ldots, x_n\right) = 1 - (1 - \ \lambda) \prod_{i=1}^{n} (1 - v_i)$$

where $\left(y = 1|x_1, \ldots, x_n\right)$ stands for the probability of the effect under the noisy-or, $\prod_{i=1}^{n} (1 - v_i)$ calculates the probability of the effect given that all causes are absent, $(1 - \ \lambda)$ specifies that also all not explicitly represented causes are absent, and finally, $1-$ takes the complement to arrive at the probability of the effect given that one or more causes are present, that is the probability of the noisy-or. When all weight parameters v_i are 1 and the leakage parameter λ is 0, then the noisy-or reduces to the Boolean OR.

The definition of the noisy-or function implies that the causes are marginally independent (such that in the absence of further information, the presence or absence of one cause does not affect the probability that other causes are present or absent) and it implies that the causes contribute independently to the effect. This means that the causal power v_i of one cause does not change with the presence or absence of other causes. In the mosquito example, one would say that the probability of contracting Ling fever from a bite of Mar remains the same whether or not we have also been bitten by Reb.

The noisy-or is the most widely used function for specifying the CPT entries in common effect structures, and experimental materials in causal reasoning research are often constructed with the aim of instantiating its independence assumptions. When these assumptions are met for a given situation or scenario, then it is possible to use the noisy-or to define the normative probability of the effect under different values of the causes. Sometimes the independence assumptions of the noisy-or have also been proposed to be descriptive of people's reasoning with common effect structures in general (Griffiths and Tenenbaum, 2009; Holyoak and Cheng, 2011), and findings of responses deviating from these assumptions have been explained as arising from people adding further information to the scenario that changes the original common effect structure into a different one (Mayrhofer et al., 2010; Rehder, 2014). In line with the default use of the noisy-or to model common effect structures, there is evidence that people find independent, additive relations between variables easier to process than interactive relations (Juslin et al., 2009; Cruz and Oberauer, 2014; Rehder and Waldmann, 2017). However, the default use of the noisy-or has also recently been criticized, partly because of concerns that it might not always be a realistic representation of causal relations in the world (Fenton et al., 2019; Noguchi et al., 2019). There can be cases in which the causes do not act independently but instead enhance or inhibit each other's contribution to the effect, and people may sometimes take account of such departures from independence in their reasoning.

This paper assesses predictions derived from the independence assumptions of the noisy-or under different conditions, and compares them to those expected under the assumption of enhancement. Inhibitory causal interaction was not considered here, but would also be worth investigating further. In the mosquito example, independent contributions of the causes to the effect can be thought of as establishing a linear relation between the number of bites from infected mosquitos and the probability of Ling fever. Causes that enhance each other's contribution to the effect could be thought of as establishing an exponential relation between number of bites and probability of Ling fever, as if once arriving in the hosts' body, the Ling bacteria coordinated their behavior to make the disease break out.

Below we discuss the predictions for independence in relation to the four conditions that result from crossing (a) whether the evidence for the effect is positive or negative, and (b) whether this evidence is certain or uncertain – and discuss how these predictions would change under the assumption of enhancement.

Condition 1: Certain Positive Evidence

Suppose we learn that the effect is present (we have Ling fever), and so increase our degree of belief in the causes (a bite from an infected mosquito of any type). If we then go on to learn that a particular cause A (e.g., a bite from Reb) is present, this "explains away" the presence of the effect. Under the noisy-or it is then normative to *decrease* again our degree of belief in the other causes (Mar and Murb). In the limit, when P(effect|cause A) = 1, cause A "explains" the presence of the effect entirely, and the probability of the other causes decreases all the way back to its baseline – the value it had before receiving the information that the effect was present. Suppose we instead go on to learn that cause A is absent. Then we are still in need of an explanation for the effect, and it is normative to augment, or increase, the probability of B. Hence under the independence assumption of the noisy-or, Condition 1 leads to the prediction of explaining away of a cause B when another cause A is present, and it leads to augmentation of a cause B when another cause A is absent.

How can the causes affect one another in this way under the noisy-or, even though they are marginally independent? When causes are *marginally independent*, then in the absence of further information, knowing that one cause is present or absent does not change the probability that another cause is present or absent. But once we learn that the effect has occurred, the causes become *conditionally dependent* on the presence of the effect. The effect establishes an indirect connection between the causes, making information about the presence or absence of one cause informative about the presence or absence of another.

Condition 2: Uncertain Positive Evidence

Suppose we do not know for sure that the effect (Ling fever) is present, but only have some uncertain indirect evidence for the effect because a consequence of the effect (e.g., a rash) is present. Then this evidence again renders the causes dependent, and it is normative under the noisy-or to show the same pattern of explaining away and augmentation as in Condition 1. The impact of uncertainty in Condition 2 is merely to decrease the size of the normative changes in probability.

Condition 3: Certain Negative Evidence

Suppose we come to know for certain that the effect is absent (we do not have Ling fever). Then it is normative to decrease our degree of belief in the causes. However, under the noisy-or the causes remain independent in this case. Additional information showing that one cause is present or absent does not undo our certainty about the absence of the effect, and so will not alter our degree of belief in the presence or absence of the other causes. Hence there is normatively no explaining away or augmentation under the noisy-or in Condition 3. It was precisely this concern about noisy-or that was addressed in Fenton et al. (2019) and Noguchi et al. (2019).

Condition 4: Uncertain Negative Evidence

Finally, suppose the effect (Ling fever) is not known for certain to be absent, but there is only some uncertain indirect evidence for this because its consequence (rash) is absent. Then the probability of the causes decreases, albeit by a smaller amount than when knowing the effect to be absent with certainty. However, because of the lingering uncertainty about whether the effect is really absent, the causes become dependent under the noisy-or. Additional information showing that one of the causes is present or absent can reduce or increase our uncertainty about the absence of the effect, and as a result, becomes informative about the probability that another cause is present or absent. Specifically, the presence of a particular cause A increases the probability of the effect, partly canceling out the reduction in the probability of the effect brought about by the absence of its consequence. As a result, the probability of an alternative cause B increases. Conversely, when A is absent, this decreases the probability of the effect, adding to the reduction in the probability of the effect brought about by the absence of its consequence. As a result, the probability of B decreases further. This pattern of probability changes goes in the opposite direction to that of explaining away and augmentation of Conditions 1 and 2.

What would follow for these four conditions if the causes did not contribute independently to the effect, but instead enhanced each other's impact? The previously described mechanisms of probability change would still be in place, but they would be overlaid by additional changes in probabilities resulting from the positive correlation between the causes. Which changes in probability prevail will depend on the relative weight of the prior probabilities and effectiveness of the causes on the one hand, and the correlation between the causes on the other.

When a positive correlation between causes is small relative to their prior probabilities and effectiveness to bring about the effect, then the direction of probability changes will be the same as for the noisy-or, although augmentation effects will be larger and explaining away effects smaller. When the prior probabilities and effectiveness of the causes are small relative to the correlation between the causes, then the impact of the correlation can override the effects predicted under independence, potentially flipping the direction of probability changes. For example, for the structure of **Figure 1**, if we know we have Ling fever and were bitten by a Reb type mosquito, then this decreases the chances that we were also bitten by a Mar type mosquito under independence. However, in a situation in which Reb and Mar very rarely bite, but when they do, they almost always bite together, then learning we were bitten by Reb might instead increase the chances that we were also bitten by Mar.

The current study did not explicitly manipulate the correlation between causes, and instead went a step back to first assess whether people's responses followed a pattern consistent with presence or absence of a correlation when this question was left open. However, the priors and effectiveness values used in this study, together with the absence of information about a potential correlation between causes, suggest it is unlikely that participants will assume a correlation between causes high enough to override the impact of priors and effectiveness information. Therefore, under the assumption of enhancement we expect explaining away to be lower in Conditions 1 and 2 than it would be under independence, but we do not expect response patterns in these conditions to flip qualitatively into augmentation. Similarly, in Condition 4 we expect the assumption of enhancement to increase the size of augmentation effects and decrease the size of explaining away effects relative to their values under independence, but we do not expect a qualitative flip from explaining away to augmentation or vice versa.

In contrast, Condition 3 does involve a qualitative difference in the predicted response patterns under assumptions of independence and of enhancement. When the effect is known to be absent with certainty, there is no explaining away or augmentation under independence. In contrast, under enhancement we expect a similar pattern of explaining away and augmentation to that predicted under the noisy-or for conditions 1 and 2, albeit again attenuated for explaining away and accentuated for augmentation. Condition 3 therefore provides a unique opportunity to differentiate whether people are interpreting causes as independent or correlated.

The above predictions are based on general principles of probability theory in a Bayesian network framework, as outlined for example in Wellman and Henrion (1993) or Morris and Larrick (1995), along with Bayesian network modeling to obtain more precise quantitative predictions for different model parameterizations (see discussion section). **Table 1** summarizes the predictions under the noisy-or for the four conditions described above.

In contrast to the extensive empirical work using noisy-or structures with positive certain evidence, there has been very little research about situations involving negative evidence, uncertain evidence, or common-effect structures that do not conform to the independence assumption of the noisy-or but instead have causes that are correlated or interact (Wellman and Henrion, 1993; Morris and Larrick, 1995; Rehder, 2014; c.f. Rottman and Hastie, 2014). For example, in one group of experiments (Rehder, 2014) participants were asked to assume that two causes contributed independently to a common effect, using relatively abstract scenarios with no information about the marginal probability of each cause. Participants were asked to compare the probability of a cause in two situations that differed in terms of whether the other cause and the effect were present, absent, or their state was unknown. When the effect was absent, participants tended to judge a given cause as equally likely regardless of the value of the other cause, as predicted by the noisy-or (case 3 above). But they also tended to judge the cause as equally likely in situations in which one would have predicted explaining away to occur. The authors explained this pattern, which is not predicted by any theory, as an aggregate of a group of participants following the predictions of the noisy-or, and another group establishing not causal but associative links between the variables involved. Associative links differ from causal links by being bidirectional rather than unidirectional. However, further research is needed

TABLE 1 | Predictions under the noisy-or for the direction of probability change of a cause B after learning that another cause A is present or absent, given four different types of evidence for the effect.

	A present	A absent
(1) Certain positive evidence	B decreases	B increases
(2) Uncertain positive evidence	B decreases slightly	B increases slightly
(3) Certain negative evidence	B remains invariant	B remains invariant
(4) Uncertain negative evidence	B increases slightly	B decreases slightly

See the general discussion for quantitative predictions for different parameterizations.

to explore alternative interpretations of these findings (see Tešić et al., 2020). Earlier studies on people's sensitivity to the impact of interactions between causes when these are made explicit in the instructions (Morris and Larrick, 1995) suggest that people's intuitions do capture the direction of the changes in probability that follow from such interactions. But the extent of such intuitions, and the contexts in which they arise, are as yet underexplored.

This paper presents an experiment intended to be a first step in assessing people's intuitions for the four conditions outlined above. To our knowledge, this is the first time that predictions under causal independence and under causal enhancement are compared directly in a single experiment, with respect to both explaining away and augmentation, and for both positive and negative evidence about the effect. The comparison also takes into account the differential impact of whether this evidence is certain or uncertain. We compared these conditions using the above mosquito scenario, which given its fictional contents was considered relatively open with respect to how the causes integrate their impact to bring about the effect. Overall, we aimed to assess which of the two integration functions accounts better for people's responses when the nature of the function is not prespecified, while taking into account that people may be uncertain about the information given even when instructed to assume it to be true or false with certainty (Evans and Over, 2004; Oaksford and Chater, 2007, 2013; Pfeifer and Kleiter, 2009; Over and Cruz, 2018).

MATERIALS AND METHODS

Participants

Fifty residents of English speaking countries completed the online experiment via the platform Prolific Academic, providing informed consent for participation. After excluding the data of participants with speeded trial responses, failed attention checks, and modest reported English language skills, the final sample consisted of 37 participants. They had a median age of 39 (range 22–65), and had a diverse formal educational background.

Materials and Design

At the start of the experiment and then again at the top of each trial, participants were shown information about a fictional archipelago in which three types of Mosquito (the Reb, Mar, and Murb mosquito) could transmit a disease known as Ling fever. The information about the mosquitos and the disease reflected the causal structure in **Figure 1**. Participants were informed that the prior probability of being bitten by an infected mosquito was 70% for each type, but that the mosquito types differed in the effectiveness with which they transmitted the disease when they bit their hosts. In the absence of bites from the other two mosquito types, the bite of an infected Reb mosquito led to Ling fever 90% of the time; the bite of an infected Mar mosquito led to Ling fever 50% of the time; and the bite of an infected Murb mosquito led to Ling fever 10% of the time. Ling fever could not be contracted through any other cause (i.e., the leakage parameter for the effect was 0). A person with Ling fever had a 90% chance of showing a purple rash. A purple rash due to other causes occurred only 10% of the time on the archipelago. The scenario made no statement about the presence or absence of any relation between causes. The above combination of parameters was chosen on the basis of a prior exploration of the parameter space in which the Bayesian network structure of **Figure 1** was queried using parameter values across the probability range, with the aim of maximizing the size of normative probability changes across conditions. The sizes of normative changes nonetheless never exceeded 20% and were sometimes smaller than 10%. We discuss the implications of this limitation further below.

The design crossed two within participant variables: (1) initial information about the effect, that is whether the effect was present (Ling fever), the effect was absent (No Ling fever), the consequence of effect was present (Rash), or the consequence of effect was absent (No rash); and (2) additional information about one of the causes (bite present vs. bite absent). Crossing these two variables resulted in eight conditions, reflected in the eight panels of **Figure 2** below.

For each of the eight conditions there were two trials, yielding 16 trials in total. On one of the trials, participants were informed that a protagonist was or was not bitten by an infected Reb mosquito, and were asked what impact this information had on the probability that the protagonist was bitten by an infected Mar mosquito. On the other trial, participants were informed that a protagonist was or was not bitten by an infected Mar mosquito, and were asked what impact this information had on the probability of them being bitten by an infected Reb mosquito. The difference between these two trials comes from the difference in effectiveness of disease transmission between mosquito types. As mentioned above, a bite from an infected Reb mosquito causes Ling fever 90% of the time, whereas a bite from an infected Mar mosquito causes Ling fever only 50% of the time. Hence, on one trial information about the presence/absence of a cause with high effectiveness is used to draw an inference about the presence/absence of a cause with medium effectiveness, and vice versa on the other trial. The Murb mosquito did not feature in the questions asked to participants because with only 10% effectiveness, this cause was associated with only very small normative changes in probability across conditions. The role of the differential effectiveness of the causes can be related to research on the reliability of testimony (Hahn et al., 2013). However, this variable goes beyond the scope of the questions addressed in this paper and its results are not discussed further here. In the context of this paper, cause effectiveness is merely a methodological variable whose inclusion makes it possible to generalize the results on the questions of interest to more than one effectiveness value. The results presented were thus averaged across the two trails for each of the eight cells of the design.

On each trial, participants were given initial information about the effect, and additional information about one of the causes. The task was to judge whether after receiving the additional information, the probability of a second cause was higher, lower, or the same compared to before receiving the additional information. The order of trials

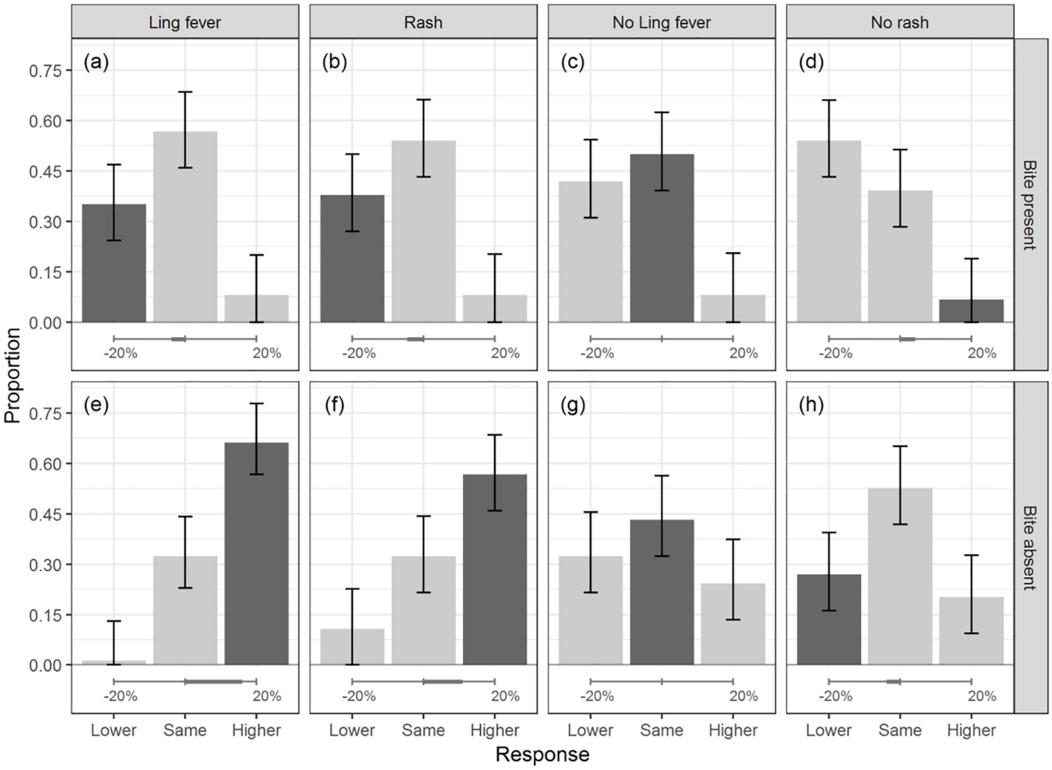


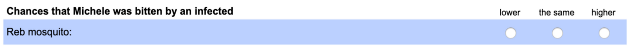
FIGURE 2 | Proportion of times each of the three response options (*lower, same, higher*) was chosen in each of the eight experimental conditions. The rows show the data separately for when the cause was present **(upper)** vs. absent **(lower)**. The columns show the data separately for the conditions in which the effect was present (Ling fever), its consequence was present (rash), the effect was absent (no Ling fever), or the consequence of the effect was absent (no rash). The dark gray bar in each panel represents the normative response under independence. The horizontal scale at the bottom of each panel shows the size of the normative change under independence. Error bars show 95% CIs.

was randomized for each participant. A screenshot of a sample trial from Condition 1 (certain positive evidence) is shown below. Each trial referred to a different island and protagonist.

Initial information for the island of Eik:

- The risk of being bitten by an infected mosquito is the same for the three mosquito types. Within a given month, a random person from the island has a 70% chance of being bitten by an infected Reb mosquito, a 70% chance of being bitten by an infected Mar mosquito, and a 70% chance of being bitten by an infected Murb mosquito.
- But the species differ in the effectiveness with which they transmit the disease when they bite their hosts.
- The Reb mosquito transmits the disease 90% of the time that it bites; the Mar mosquito 50% of the time, and the Murb mosquito 10% of the time.
- A person that has the disease has a 90% chance of showing a characteristic purple rash. The chances that a person from the island would show such a rash for other reasons is only 10%.

Michele from the island of Eik is known to have Ling fever. A further test showed that Michele was bitten by an infected Mar mosquito. Does this additional information change the chances that Michele was bitten by an infected Reb mosquito? If so, then in which way?



Chances that Michele was bitten by an infected	lower	the same	higher
Reb mosquito:	○	○	○

We asked participants to provide qualitative judgments of probability changes rather than to make repeated quantitative probability judgments under different information conditions, because we wanted to make the task less dependent on numeracy as well as on working memory limitations that could have an impact when comparing responses across trials. However, we did ask for percentage probability judgments during eight practice trials aimed at allowing participants to form an impression of the relevant causal structure and the relations between the probabilities of its elements. Two of the practice trials asked for P(cause A & cause B) and P(cause A or cause B). These probabilities allowed us to obtain an indirect impression of whether participants perceived the causes to be initially independent, that is, whether $P(A \& B) = P(A)P(B)$, and $P(A\ or\ B) = P(A) + P(B) - P(A \& B)$. We computed probabilistic coherence, that is, conformance with the axioms of probability theory, of people's responses to these two questions with and without the assumption of cause independence. This is an indirect measure because people's responses could be incoherent for many reasons (Tversky and Kahneman, 1983; Bar-Hillel and Neter, 1993). But it provides one source of information on the question, which can then be complemented with further information from this and future experiments.

Procedure

Participants went through the instructions and eight practice trials, followed by the 16 trials of the main experiment. The information on the causal structure and parameters for the scenario remained visible on each trial. At the end of the experiment, participants provided demographical information and were asked to rate on a percentage scale how difficult they found the task. The median rating of experiment difficulty was 74%. The median duration of the experimental session was 12.23 min.

RESULTS AND DISCUSSION

Coherence for participants' responses to the two practice trials asking for P(*cause A & cause B*) and P(*cause A or cause B*) was computed by first coding whether a given response was coherent or not – separately under the assumption of independence and without this assumption – and then subtracting the resulting variable for observed coherence from the chance rate of obtaining a coherent response, in order to determine whether responses were coherent more often than expected by chance (Cruz et al., 2015; Evans et al., 2015). The chance rate of a coherent response under independence is constrained to a point value as given by the equalities $P(A \& B) = P(A)P(B)$, and $P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$. In contrast, the chance rate of a coherent response without making any assumption about the relation between the causes is an interval on the probability range. For the probability of the conjunction of two causes A and B, this interval is [max(0, P(*cause A*) + P(*cause B*) − 1), min(P(*cause A*), P(*cause B*))]. For the probability of the disjuncti on of two causes it is [max(P(*cause A*), P(*cause B*)), min(P(*cause A*) + P(*cause B*), 1)].

The coherence of participants' responses to the two practice questions was found to be at chance level under the assumption of independence, but above chance when not making any assumption about how the causes might or might not be related. Specifically, assuming independence, responses to the conjunction question were coherent 7% more often than expected by chance ($t(36) = 1.56$, $p = 00.127$, 95% CI [−0.021,0.163]); and responses to the disjunction question were coherent 2% more often than expected by chance ($t(36) = 0.53$, $p = 0.533$, 95% CI [−0.038,0.072]). This outcome did not change when the range of coherent responses was increased by + −5%, and the chance rate increased accordingly, to account for the possibility that people are sensitive to the relevant coherence constraints but have degrees of belief that are coarser than point probabilities. In contrast, without assuming any specific relation between the causes, responses to the conjunction question were coherent 54% more often than expected by chance ($t(36) = 8.75$, $p < 0.001$, 95% CI [0.413,0.662]); and responses to the disjunction question were coherent 65% more often than expected by chance ($t(36) = 17.14$, $p < 0.001$, 95% CI [0.570,0.722]).

This finding does not in itself suggest that people are not assuming the causes to be independent, or that they are making no assumption about the relation between causes. But it provides an initial indication that people's probability judgments in experiments may sometimes become more understandable when moving beyond the presupposition of independence.

The pattern of responses in the main experiment is displayed in **Figure 2**. The x axis shows the three response options, and the height of the bars represents the proportion of times a response was chosen within each of the eight conditions. The darker bar in each panel shows the predicted response under independence. Each column of the figure corresponds to one of the four conditions in **Table 1**: effect present (Ling fever), consequence of effect present (rash), effect absent (no Ling fever), and consequence of effect absent (no rash). The first row represents the conditions in which an alternative cause was present, and the second row the conditions in which an alternative cause was absent. The horizontal scale at the bottom of each panel represents the size of the normative change under independence. An initial look at the figure tells us that the normative response under independence was the numerically most frequent in four of the eight experimental conditions (panels c, e, f, and g). The normative response under the assumption of a modest positive correlation between causes, whose impact is not stronger than that of the causes' priors and effectiveness values, was numerically most frequent in three of the eight conditions (panels d, e, and f).

The data were analyzed in two ways. First, a series of generalized linear models for binomial distributions compared the proportion of *higher* vs. *same*, *higher* vs. *lower*, and *same* vs. *lower* responses for each condition. A second analysis assessed, for each condition, whether the response predicted under independence was more frequent than expected by chance. This second analysis was carried out in a series of linear models following a similar procedure to the coherence analysis above. To measure whether a response was more frequent than expected by chance under independence, we first coded whether a response conformed to the prediction under independence or not, and then subtracted this variable for observed conformance from the chance rate of conforming to the predicted response. With three response options, the chance rate was 1/3 on each trial. The data were analyzed using the glm and lmer functions for the R software environment (package lme4, Bates et al. (2015); R Core Team, 2017). Analyses were performed separately for each condition because the responses predicted under independence and under enhancement changed between conditions. The general rationale for model selection aimed to maximize the random structure justified by the design, as recommended by Barr et al. (2013). However, in this case it was only possible to include random intercepts for participants in the lmer models[1].

The results show a complex picture that is not straightforward to group into findings concerning independence vs. enhancement, or explaining away vs. augmentation. We instead group the results into three domains that we think capture some of the most significant insights that can be gained from the findings, and which may explain some patterns of differences between experimental conditions.

[1] For the binomial regression analyses, effect size estimates were provided through likelihood ratio values. For the linear model analyses, which in this case were intercept only models, effect sizes were estimated in the same way as for one-sample *t*-tests: as the ratio of the fixed effect of the intercept to its standard deviation.

The Role of the Size of the Predicted Change

Consider first the conditions in which there was positive evidence for the effect (panels a, b, e, and f). Here the predictions under independence and enhancement coincide, so that any divergences from these predictions cannot easily be attributed to a violation of the assumption of causal independence.

Panels (a) and (e) show the results for when the effect (Ling fever) was certain to be present. Panels (b) and (f) show the results for when there was uncertain, indirect evidence that the effect was present because its consequence (rash) was present. Panels (e) and (f) further refer to the conditions in which one of the causes was absent. Responses in these latter two conditions showed a clear augmentation effect, in accordance with the predictions. That is, the dark bar in these panels tells us that the probability of a given cause was rated as higher upon learning that an alternative cause was absent.

Note that in these two cases the size of the predicted change under independence was larger than 10% (16.06% on average when the effect was present, and 10.95% on average when the consequence of the effect was present, as indicated through the horizontal scales at the bottom of the panels). Under enhancement, the size of the normative change would be expected to be even larger, but the extent to which it would be larger would depend on the strength of the causal interaction.

The augmentation effect was statistically significant. In panel (e) *higher* responses were more frequent than *same* responses (LR = 2.042, z = 2.865, p = 0.004, 95% CI [1.267,3.382]); and more frequent than *lower* responses (LR = 49, z = 3.853, p < 0.001, 95% CI [10.754,867.505]). The frequency of the predicted *higher* responses was above chance in this condition (*EMM* = 0.329, F(1,36) = 23.897, p < 0.001, d = 1.007, 95% CI [0.195,0.462]). In panel (f) *higher* responses were again more frequent than *same* responses (LR = 1.750, z = 2.187, p = 0.029, 95% CI [1.069,2.931]); and more frequent than *lower* responses (LR = 5.250, z = 4.299, p < 0.001, 95% CI [2.605,12.067]). The frequency of the predicted *higher* responses was also above chance in this condition (*EMM* = 0.234, F(1,36) = 11.105, p = 0.002, d = 0.690, 95% CI [0.095,0.374]).

The pattern of responses was less clear cut in panels (a) and (b). Here one of the causes is present and this "explains" the presence of the effect, leading to the prediction of a reduction in the probability of the other cause. But one can see that the size of the predicted change under independence is relatively small (3.86% on average when the effect was present, and 4.62% on average when the consequence of the effect was present, as shown in the horizontal scales at the bottom of the panels). Under enhancement, the normative size of the explaining away effect would be expected to be even smaller, albeit the extent of this decrement would again depend on the strength of the causal interaction. In line with this smaller normative change, fewer participants chose the normative *lower* response, and more participants chose the *same* response.

This pattern was corroborated statistically. For both panels (a) and (b), the frequency of the *lower* response did not differ significantly from that of *same* response (for (a): LR = 1.615, z = 1.992, p = 0.0546, 95% CI [0.997,2.666]). For (b): (LR = 1.429, z = 1.448, p = 0.148, 95% CI [0.885,2.337]); although the *lower* response was more frequent than the opposite *higher* response (For (a): LR = 0.231, z = −3.238, p = 0.001, 95% CI [0.086,0.524]. For (b): LR = 0.214, z = −3.424, p < 0.001, 95% CI [0.080,0.483]). The frequency of the *lower* response did not differ from chance in these two conditions (For (a): *EMM* = 0.018, F(1,36) = 0.067, p = 0.797, d = 0.051, 95% CI [−0.120,0.156]. For (b): *EMM* = 0.045, F(1,36) = 0.403, p = 0.529, 95%, d = 0.124, CI [−0.096,0.186]).

The pattern for panels (a) and (b) was similar to that for panel (h): the condition in which the consequence of the effect (rash) was absent and one of the causes was absent. Here the prediction under independence is that the opposite of augmentation occurs: the information that one of the causes is absent adds to the evidence for the absence of the effect, and the probability of the other cause decreases further. However, the size of the predicted change under independence was again relatively small (3.86% on average). In line with this, the *same* response was more frequent than the *lower* response (LR = 1.950, z = 2.428, p = 0.015, 95% CI [1.152,3.408]). The *lower* response was numerically more frequent than the opposite *higher* response, but this difference was not significant (LR = 0.750, z = −0.842, p = 0.400, 95% CI [0.377,1.458]). The frequency of the *lower* response was at chance level in this condition (*EMM* = −0.063, F(1,36) = 1.233, p = 0.274, d = 0.322, 95% CI [−0.176,0.050]).

The preceding results suggest people tended to respond in accordance with the probabilistic constraints given by the problem structure and in a way broadly consistent with the assumption of independence, but that differences in the frequency of relevant response options only reached significance when the normative size of the change was large enough to be noticeable (larger than 10% under independence). This was although participants made judgments only about the direction, and not about the size, of the change.

If this experiment had only tested the conditions in panels (a), (b), (e), and (f), involving positive evidence for the effect, then it would not have been possible to distinguish the role of the size of the normative change from whether this normative change was an increase (augmentation) or decrease (explaining away) of the probability of the cause asked for. That is, if we had only considered panels (a), (b), (e), and (f), then an alternative explanation for the difference in the pattern of results between (e) and (f) on the one hand, and (a) and (b) on the other, would have been that people find situations with negative evidence easier to think through than situations with positive evidence. But such an alternative explanation does not fit with the results for panel (h), which concern negative evidence and yet resemble the responses given to the cases of positive evidence in (a) and (b) more than those for negative evidence in (e) and (f). Considering the five panels together, a better, and simpler, explanation for the differences between conditions seems to be that they reflect differences in the size of the normative change. Further experiments varying the size of the normative change systematically across conditions would be necessary to further test this interpretation.

The Probability of a Cause When the Effect Is Absent

Let us now turn to panels (c) and (g): the conditions in which the effect (Ling fever) was certain to be absent. Under the assumption of causal independence, the normative response in these two conditions is that there is no change. If the causes are instead positively correlated, then the normative response is explaining away for (c) and augmentation for (g). Finally, if the causes are interpreted as contributing independently to the effect but the absence of the effect is treated as uncertain (Oaksford and Chater, 2013; Over and Cruz, 2018), then the normative response is the opposite of explaining away and augmentation: an increase in the probability of a cause when another cause is present (c), and a decrease in the probability of a cause when another cause is absent (g). The conditions of panels (c) and (g) offer a unique opportunity for testing the contrasting predictions of the above three assumptions.

Consider first the condition in panel (c), where one of the causes is present. The *same* response predicted under independence was numerically more frequent than the *lower* response predicted under enhancement, but the difference was not significant (LR = 1.194, z = 0.727, p = 0.467, 95% CI [0.741,1.934]). The *same* and the *lower* response were both more frequent than the *higher* response predicted under the assumption of independence + uncertainty (*same* vs. *higher*: LR = 0.162, z = −4.133, p < 0.001, 95% CI [0.062,0.356]. *Lower* vs. *higher*: LR = 0.194, z = −3.682, p < 0.001, 95% CI [0.073,0.432]). The *same* response predicted under independence was above chance in this condition (*EMM* = 0.167, F(1,36) = 4.933, p = 0.033, d = 0.415, 95% CI [0.018,0.316]). Overall, the responses in this condition were in accordance with independence and, to a numerically lesser extent, with enhancement. In contrast, there was no evidence that participants followed the independence assumption while treating the information that the effect was absent as uncertain.

Turning to the condition in panel (g), where one of the causes is absent, the numerically most frequent response was again that there is no change, in line with the independence assumption. But the pattern was less clear cut, and no response option was significantly more frequent than the others (*same* vs. *lower*: LR = 1.333, z = 1.065, p = 0.287, 95% CI [0.788,2.286]. *Same* vs. *higher*: LR = 0.563, z = −1.953, p = 0.051, 95% CI [0.310,0.991]. *Lower* vs. *higher*: LR = 0.750, z = −0.923, p = 0.356, 95% CI [0.401,1.376]). The frequency of the *same* response was at chance level in this condition (*EMM* = 0.099, F(1,36) = 2.151, p = 0.151, 95%, d = 0.334, CI [−0.035,0.233]). Overall, participants seemed to have no clear common intuitions for the case in which both the effect and one of the causes was absent.

Evidence Against Negative Evidence

Finally, consider the pattern in panel (d). Here there is uncertain evidence that the effect (Ling fever) is absent because its consequence (rash) is absent, and we then learn that one of the causes is present. For the parameters of the model, the predicted response under both independence and enhancement assumptions is that the opposite of explaining away occurs. This is because the presence of the cause undermines the uncertain evidence for the absence of the effect. The probability of the effect increases again, and with it also the probability of the other cause. The predicted size of the change under independence was 4.14% on average, which is not very large. Considering the findings from panels (a), (b), and (h), we can thus expect a relatively high frequency of same responses in this condition. The panel shows that although there was indeed a sizeable number of same responses, the most frequent response was instead the *lower* response, which is opposite to what had been predicted.

Statistically, the predicted *higher* responses were less frequent than the *same* responses (LR = 0.172, z = −3.630, p < 0.001, 95% CI [0.059,0.408]) and less frequent than the *lower* responses (LR = 0.125, z = −4.384, p < 0.001, 95% CI [0.043,0.288]). The frequency of the *higher* response was below chance in this condition (*EMM* = −0.266, F(1,36) = 59.504, p < 0.001, d = 1.728, 95% CI [−0.334,−0.197]).

The finding for this condition was surprising, and is the only one of the eight investigated in which responses seemed to deviate systematically from Bayesian predictions under both independence and enhancement assumptions. One possible explanation is that it constitutes a double negation effect. This effect, first described in research on deductive reasoning, refers to the finding that people make more errors drawing inferences when this requires negating a negation. That is, when it requires establishing that *not-not-A* = *A* (Evans and Handley, 1999; Oaksford et al., 2000). In a probabilistic extension of this idea, the present condition required participants to undermine negative evidence for an effect, and assess the consequences of this for the probability of a cause. However, the finding would have to be replicated and the conditions of its occurrence investigated further to determine the value of this explanation.

DISCUSSION

This study investigated people's intercausal judgments in situations with several alternative causes for a common effect. We compared the predictions that follow from assuming that the causes contribute independently to the effect, with those that follow from assuming that the causes interact to some extent, enhancing each other's contribution to the effect. In doing so, we took into account: (a) whether the information about the effect was considered certain or uncertain, (b) whether the evidence for the effect was positive or negative, and (c) whether one of the causes was present or absent. The resulting eight conditions were compared in a single experiment using a within participants design.

The experiment aimed to explore further people's intuitions about explaining away and augmentation, and identify possible factors that could shed light on why previous studies have often found people's responses to conform with the explaining away effects that follow from the independence assumption of the noisy-or, but to a lesser extent than predicted by normative models (Ali et al., 2011; Rehder, 2014; Rottman and Hastie, 2014).

Extant explanations for under-explaining away have pointed to possible differences in the interpretation of probability (Tešić et al., 2020), prior knowledge that changes the causal structure reasoned about, for instance by adding links between causes or additional intervening variables that must be active to allow an effect to occur (Mayrhofer et al., 2010; Rehder, 2014; Rottman and Hastie, 2014), and by positing that a subset of participants may represent the relations between variables as associative, and thus bidirectional, rather than as causal and unidirectional (Rehder and Waldmann, 2017). The latter has been referred to as the "rich get richer" principle because it implies that when one variable is present, this will increase the probability that variables connected to it will also be present, and vice-versa when a variable is absent (c.f. parallel constraint satisfaction networks, Glöckner et al., 2010).

The present results highlight two further possible reasons for the findings of under-explaining away, which we view as complementing rather than standing in competition to the explanations outlined above. The first is that people may not spontaneously interpret causes as contributing independently to the effect, as presupposed by the use of the noisy-or, but may sometimes instead interpret the causes as enhancing each other's contribution, even in cases in which the materials are fictional and no explicit information suggesting any relation between causes is provided. On the one hand, this underlines the need to be careful when designing experiments, to make sure participants are really assuming causal independence before interpreting deviations from the predictions under independence as non-normative. On the other hand, it also points to the option of not trying to create materials for which independence unambiguously holds in the first place, and instead setting out to examine in more detail how people reason about causal structures with interacting causes.

The absence of a manipulation of the size of a correlation or interaction between causes was a limitation of the current study, and something worth pursuing in follow-up work. Such work could also include a dissociation of the two independence assumptions of the noisy-or separately, exploring separately people's intuitions about (a) causes that covary, in the sense that the marginal probability that one cause is present changes as a function of the probability of another cause (Rottman and Hastie, 2014); and (b) causes that interact in their contribution to the effect, in the sense that whenever two causes happen to be present at the same time, the probability of the effect is increased or decreased to a greater extent than would be predicted by considering the impact of each cause independently (see also Fenton et al., 2019). An example of covariance would be a situation in which one cook is preparing soup, and the smell of the soup compels other cooks to enter the kitchen and start cooking more soup. An example of interaction would be a situation in which whenever soup happens to be cooked by more than one cook, the cooks start to work together; making the soup turn out better/worse than it would have been if they had been working independently.

Studies of reasoning about covarying and interacting causes are made more difficult by the lack of a single function from which to derive the CPT for the causal structure of interest. But this difficulty can be met by determining the size of each interaction effect, and then modifying an initial CPT based on independence to incorporate the interaction (Wellman and Henrion, 1993; Lemmer and Gossink, 2004; Fenton et al., 2019).

A second possible reason for the under-explaining away found in previous studies is that at least in some of these studies, the size of the normative change itself may have been too small to be subjectively relevant. Thus, people might be sensible to the probabilistic constraints posed by the structure of the problem, but our degrees of belief may be coarser than point probabilities, so that a larger change is necessary for it to be subjectively meaningful. The size of the normative change is not available in studies that do not include precise information about priors and causal power, and this information is of course often not available in real world situations (Rehder, 2014; Rottman and Hastie, 2014). However, the present findings suggest that in those cases in which the size of the normative change is not negligible, people's responses do follow normative predictions in a consistent way.

As a further argument for the above interpretation, **Figure 3** shows the size of the normative change that occurs under the assumption of independence for a causal structure like that of **Figure 1**. As in **Figure 1**, the leakage parameter for the consequence of the effect (rash) was set to 0.1, but unlike **Figure 1**, the causes were set to have equal priors and equal causal power for simplicity.

Figure 3 shows the size of the normative change for three of the four conditions of **Table 1**. The fourth condition: when the effect is absent, was not included in **Figure 3** because it is associated with the prediction that the probabilities of the causes remain invariant, i.e., there is no normative change in this case.

The left column shows the condition in which the effect is known to be present – the most commonly studied case for explaining away in the literature. The middle column shows the condition in which the consequence of the effect is absent, and the right column shows the condition in which the consequent of the effect is present. For each condition, the size of the normative change is shown on the y axis as a function of the prior probabilities of the causes (on the x axis), the causal power of the causes (separate lines) and the leakage parameter λ (separate rows).

One can see that across the range of values these parameters can take, the size of the normative change only rarely reaches values higher than 20%, and it decreases as the value of the leakage parameter increases. When the effect is present or its consequence is absent, the size of the change tends to be larger for lower values of the prior, whereas this is not the case when the consequence of the effect is present. Overall, the size of the normative change increases with the power of the causes. When the effect is present, this effect of causal power is more or less evenly spread. In contrast, when the consequence is absent, it takes a causal power over 0.5 to obtain a non-negligible change at all, with higher values of causal power having increasing impact.

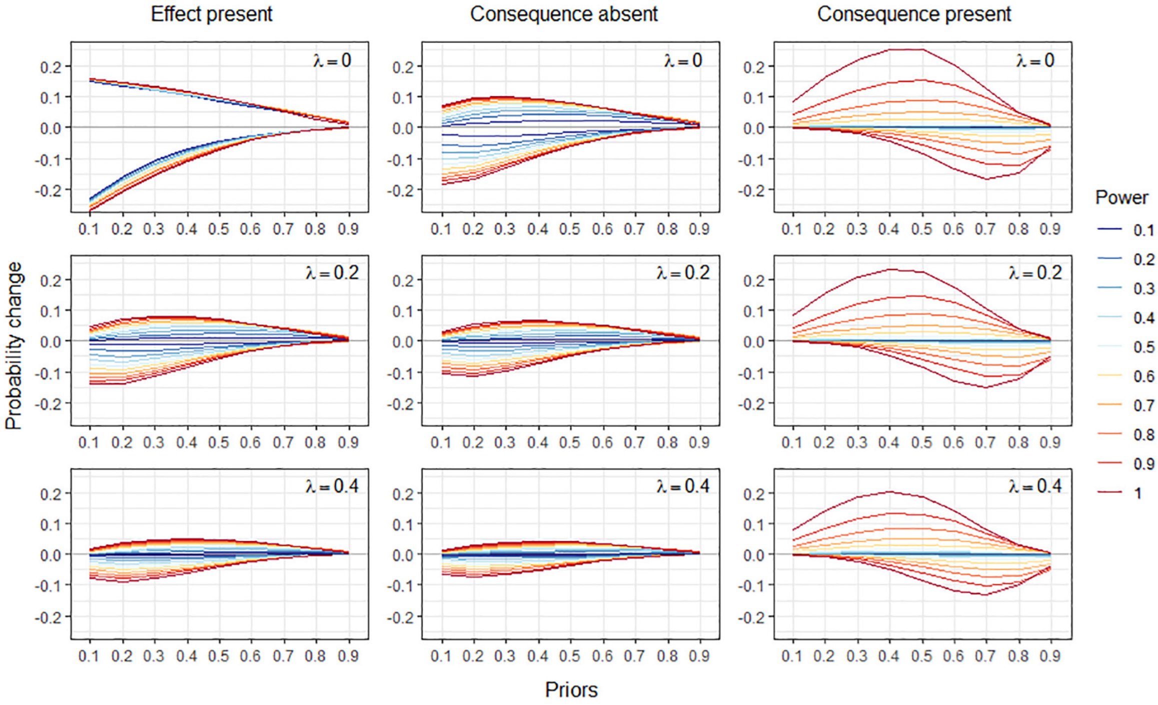


FIGURE 3 | Normative changes in the probability of one cause when learning that another cause is present or absent, for a common-effect structure like that of **Figure 1** but with equal priors and causal powers for each cause. Probability changes are shown as a function of the prior probabilities of the causes (x axis), causal power (separate lines), and the value of the leakage parameter (separate rows). Left column: explaining away and augmentation for the condition in which the effect is present. Middle column: opposite probability changes to those of explaining away and augmentation for the condition in which the consequence of the effect is absent. Right column: explaining away and augmentation for the condition in which the consequence of the effect is present. The condition in which the effect is absent is not shown here because probabilities remain invariant under the noisy-or in this case (c.f. **Table 1**).

Note that as mentioned above, positive and negative normative changes arise under opposite conditions when the effect is present and when the consequence of the effect is absent. When the effect is present, negative changes correspond to the size of the explaining away effect, P(*cause B*| *effect & cause A*) - P(*cause B*| *effect*) < 0, and positive changes correspond to the size of the augmentation effect, P(*cause B*| *effect & not-cause A*) - P(*cause B*| *effect*) > 0. In contrast, when the consequence is absent, the normative probability changes go in the opposite direction: P(*cause B*| *effect & cause A*) - P(*cause B*| *effect*) > 0, and P(*cause B*| *effect & not-cause A*) - P(*cause B*| *effect*) < 0. This distinction is not visible in the graphs, which focus instead on illustrating the impact of causal power.

In the present study it was difficult to find model parameters for which the size of the predicted change was substantial across all experimental conditions in which a change was predicted, which included conditions in which the evidence for the effect was negative. But further studies could test this factor more explicitly by varying the size of the normative change within each condition and assessing the effect of this variation on participants' probability judgments.

Future work could also assess the generalizability of the present findings by asking participants for numeric probability judgments under different information conditions in a dynamic reasoning setting, rather than for qualitative probability changes as was done here. This would also make it easier to build up the task more gradually for participants, for instance asking first about P(*effect*), then about P(*effect*| *not-cause A*), and finally about P(*effect*| *cause B & not-cause A*).

Finally, we found that responses were contrary to predictions under both independence and enhancement assumptions when a partial canceling of the effect of negative evidence was required. This unexpected finding resembles a probabilistic extension of the double negation effect from the deductive reasoning literature, and is worth investigating further. If replicated it may constitute a distinct source of error in probabilistic reasoning, beyond more frequently discussed sources such as those based on content effects, associative reasoning, and the use of heuristic task simplifications (Tversky and Kahneman, 1983; Oaksford, 2002; Glöckner et al., 2010; Rehder and Waldmann, 2017).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Department of Psychological Sciences of Birkbeck, University of London. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

NC implemented the study, analyzed the data, and wrote the manuscript. All authors contributed to informing the theoretical background and hypotheses and manuscript revision.

FUNDING

This research was based upon work supported in part by ERC project ERC-2013-AdG339182-BAYES_KNOWLEDGE, Leverhulme Trust grant RPG-2016-118 CAUSAL-DYNAMICS, ESRC grant ES/H041796/1 Causality and responsibility judgments, the Alexander von Humboldt Foundation, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), to the BARD project (Bayesian Reasoning via Delphi) of the CREATE programme under Contract (2017-16122000003). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government was authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

Ali, N., Chater, N., and Oaksford, M. (2011). The mental representation of causal conditional reasoning: mental models or causal models. *Cognition* 119, 403–418. doi: 10.1016/j.cognition.2011.02.005

Bar-Hillel, M., and Neter, E. (1993). How alike is it versus how likely is it: a disjunction fallacy in probability judgments. *J. Personal. Soc. Psychol.* 65, 1119–1131. doi: 10.1037/0022-3514.65.6.1119

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., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv:1506.04967v1 [stat.ME]*. Available online at: http://arxiv.org/abs/1506.04967

Briggs, R. (2012). Interventionist counterfactuals. *Phil. Stud.* 160, 139–166. doi: 10.1007/s11098-012-9908-5

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychol. Rev.* 104, 367–405. doi: 10.1037/0033-295x.104.2.367

Cruz, N., Baratgin, J., Oaksford, M., and Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Front. Psychol.* 6:192. doi: 10.3389/fpsyg.2015.00192

Cruz, N., and Oberauer, K. (2014). Comparing the meanings of "if" and "all". *Mem. Cogn.* 42, 1345–1356. doi: 10.3758/s13421-014-0442-x

Evans, J. St. B. T, and Handley, S. J. (1999). The role of negation in conditional inference. *Q. J. Exp. Psychol.* 52A, 739–769. doi: 10.1080/713755834

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

Evans, J. St. B. T, Thompson, V., and Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Front. Psychol.* 6:398. doi: 10.3389/fpsyg.2015.00398

Fenton, N., Neil, M., and Caballero, J. G. (2007). Using ranked nodes to model qualitative judgments in Bayesian networks. *IEEE Trans. Knowledge Data Eng.* 19, 1420–1432. doi: 10.1109/tkde.2007.1073

Fenton, N., Noguchi, T., and Neil, M. (2019). An extension to the noisy-OR function to resolve the 'explaining away' deficiency for practical Bayesian network problems. *IEEE Trans. Knowledge Data Eng.* 31, 2441–2445. doi: 10.1109/tkde.2019.2891680

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

Glöckner, A., Betsch, T., and Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *J. Behav. Dec. Mak.* 23, 439–462. doi: 10.1002/bdm.668

Griffiths, T. L., and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychol. Rev.* 116, 661–716. doi: 10.1037/a0017201

Hahn, U., Oaksford, M., and Harris, A. J. L. (2013). "Testimony and argument: a bayesian perspective," in *Bayesian Argumentation: the Practical side of Probability*, ed. F. Zenker (New York, NY: Springer).

Holyoak, K. J., and Cheng, P. W. (2011). Causal learning and inference as a rational process: the new synthesis. *Ann. Rev. Psychol.* 62, 135–163. doi: 10.1146/annurev.psych.121208.131634

Juslin, P., Nilsson, H., and Winman, A. (2009). Probability theory, not the very guide of life. *Psychol. Rev.* 116, 856–874. doi: 10.1037/a0016979

Lemmer, J. F., and Gossink, D. E. (2004). Recursive noisy OR – a rule for eetimating complex probabilistic interactions. *IEEE Trans. Syst. Man Cybern. Part B* 34, 2252–2261. doi: 10.1109/tsmcb.2004.834424

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Liefgreen, A., Tešiæ, M., and Lagnado, D. (2018). "Explaining away: significance of priors, diagnostic reasoning, and structural complexity," in *The 40th Annual Meeting of the Cognitive Science Society*, eds C. Kalish, M. Rau, J. Xhu, and T. T. Rogers (Madison WI: Cognitive Science Society). Chairs.

Mayrhofer, R., Hagmayer, Y., and Waldmann, M. R. (2010). "Agents and causes: a bayesian error attribution model of causal reasoning," in *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society*, eds S. Ohlsson and R. Catrambone (Austin, Tex: Cognitive Science Society).

Morris, M. W., and Larrick, R. P. (1995). When one cause casts doubt on another: a normative analysis of discounting in causal attribution. *Psychol. Rev.* 102, 331–355. doi: 10.1037/0033-295x.102.2.331

Noguchi, T., Fenton, N. E., and Neil, M. (2019). Addressing the practical limitations of Noisy-OR using conditional inter-causal anti-correlation with ranked nodes. *IEEE Trans. Knowledge Data Eng.* 31, 813–817. doi: 10.1109/tkde.2018.2873314

Oaksford, M. (2002). Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Think. Reason.* 8, 135–151. doi: 10.1080/13546780143000170

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

Oaksford, M., and Chater, N. (2017). "Causal models and conditional reasoning," in *The Oxford Handbook of Causal Reasoning*, ed. M. Waldmann (Oxford: Oxford University Press).

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

Over, D. E. (2017). "Causation and the probability of causal conditionals," in *The Oxford Handbook of Causal Reasoning*, ed. M. Waldmann (Oxford: Oxford University Press).

Over, D. E., and Cruz, N. (2018). "Probabilistic accounts of conditional reasoning," in *International Handbook of thinking and reasoning*. eds L. J. Ball and V. A. Thompson (Hove: Psychology Press).

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan-Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York, NY: Cambridge University Press.

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

R Core Team. (2017). *R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna: R Core Team.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cogn. Psychol.* 72, 54–107. doi: 10.1016/j.cogpsych.2014.02.002

Rehder, B., and Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Mem. Cogn.* 45, 245–260. doi: 10.3758/s13421-016-0662-3

Rottman, B. M., and Hastie, R. (2014). Reasoning about causal relationships: inferences on causal networks. *Psychol. Bull.* 140, 109–139. doi: 10.1037/a0031903

Sloman, S. A., and Lagnado, D. (2005). Do we "do"? *Cogn. Sci.* 29, 5–39.

Sloman, S. A., and Lagnado, D. (2015). Causality in thought. *Ann. Rev. Psychol.* 66, 223–247.

Stalnaker, R. (1981). "A theory of conditionals," in *Ifs*, eds W. Harper, R. Stalnaker, and G. Pearce (Dordrecht: D. Reidel). doi: 10.1075/cilt.143.12ziv

Tešić, M., Liefgreen, A., and Lagnado, D. (2020). The propensity interpretation of probability and diagnostic split in explaining away. *Cogn. Psychol.* 121:101293. doi: 10.1016/j.cogpsych.2020.101293

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

Waldmann, M. R., and Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *J. Exp. Psychol. General* 121, 222–236. doi: 10.1037/0096-3445.121.2.222

Wellman, M. P., and Henrion, M. (1993). Explaining "Explaining away". *IEEE Trans. Pattern Anal. Mach. Intell.* 15, 287–292.

Conflict of Interest: 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 reviewer AT declared a shared affiliation with one of the authors DL to the handling editor at the time of review.

frontiers in Psychology

REVIEW
published: 05 January 2021
doi: 10.3389/fpsyg.2020.513474

Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It

*J. Mark Bishop**

Department of Computing, Goldsmiths, University of London, London, United Kingdom

Artificial Neural Networks have reached "grandmaster" and even "super-human" performance across a variety of games, from those involving perfect information, such as Go, to those involving imperfect information, such as "Starcraft". Such technological developments from artificial intelligence (AI) labs have ushered concomitant applications across the world of business, where an "AI" brand-tag is quickly becoming ubiquitous. A corollary of such widespread commercial deployment is that when AI gets things wrong—an autonomous vehicle crashes, a chatbot exhibits "racist" behavior, automated credit-scoring processes "discriminate" on gender, etc.—there are often significant financial, legal, and brand consequences, and the incident becomes major news. As Judea Pearl sees it, the underlying reason for such mistakes is that "*... all the impressive achievements of deep learning amount to just curve fitting*." The key, as Pearl suggests, is to replace "reasoning by association" with "causal reasoning" —the ability to infer causes from observed phenomena. It is a point that was echoed by Gary Marcus and Ernest Davis in a recent piece for the *New York Times*: "*we need to stop building computer systems that merely get better and better at detecting statistical patterns in data sets—often using an approach known as 'Deep Learning'—and start building computer systems that from the moment of their assembly innately grasp three basic concepts: time, space, and causality*." In this paper, foregrounding what in 1949 Gilbert Ryle termed "a category mistake", I will offer an alternative explanation for AI errors; it is not so much that AI machinery cannot "grasp" causality, but that AI machinery (qua computation) cannot understand anything at all.

Keywords: dancing with pixies, Penrose-Lucas argument, causal cognition, artificial neural networks, artificial intelligence, cognitive science, Chinese room argument

OPEN ACCESS

Edited by:
*Andrew Tolmie,
University College London,
United Kingdom*

Reviewed by:
*Manuel Bedia,
University of Zaragoza, Spain
Wolfgang Schoppek,
University of Bayreuth, Germany*

***Correspondence:**
*J. Mark Bishop
m.bishop@gold.ac.uk*

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

Received: 20 November 2019
Accepted: 07 September 2020
Published: 05 January 2021

Citation:
*Bishop JM (2021) Artificial Intelligence
Is Stupid and Causal Reasoning Will
Not Fix It. Front. Psychol. 11:513474.
doi: 10.3389/fpsyg.2020.513474*

1. MAKING A MIND

For much of the twentieth century, the dominant cognitive paradigm identified the mind with the brain; as the Nobel laureate Francis Crick eloquently summarized:

> "You, your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules. As Lewis Carroll's Alice might have phrased, 'You're nothing but a pack of neurons'. This hypothesis is so alien to the ideas of most people today that it can truly be called astonishing" (Crick, 1994).

Motivation for the belief that a computational simulation of the mind is possible stemmed initially from the work of Turing (1937) and Church (1936) and the "Church-Turing hypothesis"; in Turing's formulation, every "function which would naturally be regarded as computable" can be computed by the "Universal Turing Machine." If computers can adequately model the brain, then, theory goes, it ought to be possible to *program* them to act like minds. As a consequence, in the latter part of the twentieth century, Crick's "Astonishing Hypothesis" helped fuel an explosion of interest in connectionism: both high-fidelity simulations of the brain (computational neuroscience; theoretical neurobiology) and looser—merely "neural inspired" —analoges (cf. Artificial Neural Networks, Multi-Layer Perceptrons, and "Deep Learning" systems).

But the fundamental question that Crick's hypothesis raises is, of course, that if we ever succeed in fully instantiating a *sufficiently accurate* simulation of the brain on a digital computer, will we also have fully instantiated a digital [computational] mind, with all the human mind's causal power of teleology, understanding, and reasoning, and will artificial intelligence (AI) finally have succeeded in delivering "Strong AI"[1].

Of course, *if* strong AI is possible, accelerating progress in its underpinning technologies[2]–entailed both by the use of AI systems to design ever more sophisticated AIs and the continued doubling of raw computational power every 2 years[3]—will eventually cause a runaway effect whereby the AI will inexorably come to exceed human performance on all tasks[4]; the so-called point of [technological] "singularity" ([in]famously predicted by Ray Kurzweil to occur as soon as 2045[5]). And, at the point this "singularity" occurs, so commentators like Kevin Warwick[6] and Stephen Hawking[7] suggest, humanity will, effectively, have been "superseded" on the evolutionary ladder and be obliged to eke out its autumn days listening to "Industrial Metal" music and gardening; or, in some of Hollywood's even more dystopian dreams, cruelly subjugated (and/or exterminated) by "Terminator" machines.

In this paper, however, I will offer a few "critical reflections" on one of the central, albeit awkward, questions of AI: why is it that, seven decades since Alan Turing first deployed an "effective method" to play chess in 1948, we have seen enormous strides in engineering particular machines to do clever things—from driving a car to beating the best at Go—but almost no progress in getting machines to genuinely understand; to seamlessly apply knowledge from one domain into another—the so-called problem of "Artificial General Intelligence" (AGI); the skills that both Hollywood and the wider media really think of, and depict, as AI?

2. NEURAL COMPUTING

The earliest cybernetic work in the burgeoning field of "neural computing" lay in various attempts to understand, model, and emulate neurological function and learning in animal brains, the foundations of which were laid in 1943 by the neurophysiologist Warren McCulloch and the mathematician Walter Pitts (McCulloch and Pitts, 1943).

Neural Computing defines a mode of problem solving based on "learning from experience" as opposed to classical, syntactically specified, "algorithmic" methods; at its core is "*the study of networks of 'adaptable nodes' which, through a process of learning from task examples, store experiential knowledge and make it available for use*" (Aleksander and Morton, 1995). So construed, an "Artificial Neural Network" (ANN) is constructed merely by appropriately connecting a group of adaptable nodes ("artificial neurons").

- A *single layer neural network* only has one layer of adaptable nodes between the input vector, X and the output vector O, such that the output of each of the adaptable nodes defines one element of the network output vector O.
- A *multi-layer neural network* has one or more "hidden layers" of adaptable nodes between the input vector and the network output; in each of the network *hidden layers*, the outputs of the adaptable nodes connect to one or more inputs of the nodes in subsequent layers and in the network *output layer*, the output of each of the adaptable nodes defines one element of the network output vector O.
- A *recurrent neural network* is a network where the output of one or more nodes is fed-back to the input of other nodes in the architecture, such that the connections between nodes form a "directed graph along a temporal sequence," so enabling a recurrent network to exhibit "temporal dynamics," enabling a recurrent network to be sensitive to particular *sequences* of input vectors.

[1] Strong AI, a term coined by Searle (1980) in the "Chinese room argument" (CRA), entails that, "*... the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states*," which Searle contrasted with "Weak AI" wherein "*... the principal value of the computer in the study of the mind is that it gives us a very powerful tool.*" Weak AI focuses on epistemic issues relating to engineering a simulation of [human] intelligent behavior, whereas strong AI, in seeking to engineer a computational system with all the causal power of a mind, focuses on the ontological.

[2] See "[A]mplifiers for intelligence—devices that supplied with a little intelligence will emit a lot" (Ashby, 1956).

[3] See Moore's law: the observation that the number of transistors in a dense integrated circuit approximately doubles every 2 years.

[4] Conversely, as Francois Chollet, a senior engineer at Google and well-known scptic of the "Intelligence Explosion" scenario; trenchantly observed in 2017: "*The thing with recursive self-improvement in AI, is that if it were going to happen, it would already be happening. Auto-Machine Learning systems would come up with increasingly better Auto-Machine Learning systems, Genetic Programming would discover increasingly refined GP algorithms*" and yet, as Chollet insists, "*no human, nor any intelligent entity that we know of, has ever designed anything smarter than itself.*"

[5] Kurzweil (2005) "set the date for the Singularity—*representing a profound and disruptive transformation in human capability*—as 2045."

[6] In his 1997 book "March of the Machines", Warwick (1997) observed that there were already robots with the "*brain power of an insect*"; soon, or so he predicted, there would be robots with the "*brain power of a cat*," and soon after that there would be "*machines as intelligent as humans.*" When this happens, Warwick darkly forewarned, the science-fiction nightmare of a "Terminator" machine could quickly become reality because such robots will rapidly, and inevitably, become more intelligent and superior in their practical skills than the humans who designed and constructed them.

[7] In a television interview with Professor Stephen Hawking on December 2nd 2014, Rory Cellan-Jones asked how far engineers had come along the path toward creating Artificial Intelligence, to which Professor Hawking alarmingly replied, "*Once humans develop artificial intelligence it would take off on its own and redesign itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded.*"

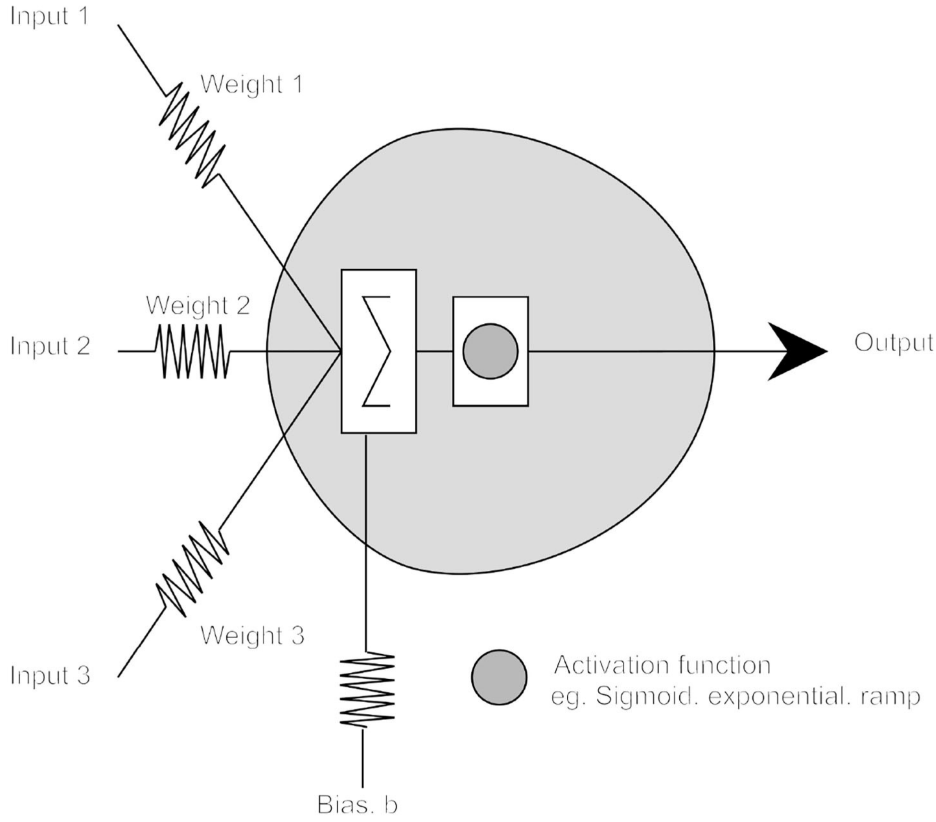


FIGURE 1 | The McCulloch–Pitts neuron model.

Since 1943 a variety of frameworks for the adaptable nodes have been proposed[8]; however, the most common, as deployed in many "deep" neural networks, remains grounded on the McCulloch/Pitts model.

2.1. The McCulloch/Pitts (MCP) Model

In order to describe how the basic processing elements of the brain might function, McCulloch and Pitts showed how simple electrical circuits, connecting groups of "linear threshold functions," could compute a variety of logical functions (McCulloch and Pitts, 1943). In their model, McCulloch and Pitts provided a first (albeit very simplified) mathematical account of the chemical processes that define neuronal operation and in so doing realized that the mathematics that describe the neuron operation exhibited exactly the same type of logic that Shannon deployed in describing the behavior of switching circuits: namely, the calculus of propositions.

McCulloch and Pitts (1943) realized (a) that neurons can receive positive or negative encouragement to fire, contingent upon the type of their "synaptic connections" (excitatory or inhibitory) and (b) that in firing the neuron has effectively performed a "computation"; once the effect of the excitatory/inhibitory synapses are taken into account, it is possible to *arithmetically* determine the net effect of incoming patterns of "signals" innervating each neuron.

In a simple McCulloch/Pitts (MCP) threshold model, adaptability comes from representing each synaptic junction by a variable (usually rational) valued weight W_i, indicating the degree to which the neuron should react to the $_ith$ particular input (see **Figure 1**). By convention, positive weights represent excitatory synapses and negative, inhibitory synapses; the neuron firing threshold being represented by a variable T. In modern use, T is usually clamped to zero and a threshold implemented using a variable "bias" weight, b; typically, a neuron firing[9] is represented by the value +1 and not firing by 0.

Activity at the i_{th} input to an n input neuron is represented by the symbol X_i and the effect of the i_{th} synapse by a weight W_i, hence the net effect of the i_{th} input on the i_{th} synapse on the MCP

[8] These include "spiking neurons" as widely used in computational neuroscience (Hodgkin and Huxley, 1952); "kernel functions" as deployed in "Radial Basis Function" networks (Broomhead and Lowe, 1988) and "Support Vector Machines" (Boser et al., 1992); "Gated MCP Cells," as deployed in LSTM networks (Hochreiter and Schmidhuber, 1997); "n-tuple" or "RAM" neurons, as used in "Weightless" neural network architectures (Bledsoe and Browning, 1959; Aleksander and Stonham, 1979), and "Stochastic Diffusion Processes" (Bishop, 1989) as deployed in the NESTOR multi-variate connectionist framework (Nasuto et al., 2009).

[9] "In psychology.. the fundamental relations are those of two valued logic" and McCulloch and Pitts recognized neuronal firing as equivalent to "representing" a proposition as *TRUE* or *FALSE* (McCulloch and Pitts, 1943).

cell is thus $X_i \times W_i$. Thus, the MCP cell is denoted as firing if:

$$\sum_{i}^{n} X_i \times W_i + b \geq 0 \tag{1}$$

In a subsequent generalization of the basic MCP neuron, cell output is defined by a further (typically non-linear) function of the weighted sum of its input, the neuron's *activation function*.

McCulloch and Pitts (1943) proved that if "synapse polarity" is chosen appropriately, any single pattern of input can be "recognized" by a suitable network of MCP neurons (i.e., any finite logical expression can be realized by a suitable network of McCulloch–Pitts neurons). In other words, the McCulloch–Pitts' result demonstrated that networks of artificial neurons could be mathematically specified, which would perform "computations" of immense complexity and power and in so doing, opened the door to a form of problem solving based on the design of appropriate neural network architectures and automatic (machine) "learning" of appropriate network parameters.

3. EMBEDDINGS IN EUCLIDEAN SPACE

The most commonly used framework for information representation and processing in artificial neural networks (via generalized McCulloch/Pitts neurons) is a subspace of Euclidean space. Supervised learning in this framework is equivalent to deriving appropriate transformations (learning appropriate mappings) from training data (problem exemplars; pairs of *Input* + "*Target Output*″ vectors). The majority of learning algorithms adjust neuron interconnection weights according to a specified "learning rule," the adjustment in a given time step being a function of a particular training example.

Weight updates are successively aggregated in this manner until the network reaches an equilibrium, at which point no further adjustments are made or, alternatively, learning stops before equilibrium to avoid "overfitting" the training data. On completion of these computations, knowledge about the training set is represented across a distribution of final weight values; thus, a trained network does not possess any internal representation of the (potentially complex) relationships *between* particular training exemplars.

Classical multi-layer neural networks are capable of discovering non-linear, continuous transformations between objects or events, but nevertheless they are restricted by operating on representations embedded in the linear, continuous structure of Euclidean space. It is, however, doubtful whether regression constitutes a satisfactory (or the most general) model of information processing in natural systems.

As Nasuto et al. (1998) observed, the world, and relationships between objects in it, is fundamentally non-linear; relationships between real-world objects (or events) are typically far too messy and complex for representations in Euclidean spaces—and smooth mappings between them—to be appropriate embeddings (e.g., entities and objects in the real-world are often fundamentally discrete or qualitatively vague in nature, in which case Euclidean space does not offer an appropriate embedding for their representation).

Furthermore, representing objects in a Euclidean space imposes a serious additional effect, because Euclidean vectors can be compared to each other by means of *metrics*; enabling data to be compared in spite of any real-life constraints (sensu stricto, metric rankings may be undefined for objects and relations of the real world). As Nasuto et al. (1998) highlight, it is not usually the case that all objects in the world can be equipped with a "natural ordering relation"; after all, what is the natural ordering of "banana" and "door"?

It thus follows that classical neural networks are best equipped only for tasks in which they process numerical data whose relationships can be reflected by Euclidean distance. In other words, classical connectionism can be reasonably well-applied to the same category of problems, which could be dealt with by various regression methods from statistics; as Francois Chollet[10], in reflecting on the limitations of deep learning, recently remarked:

> "[a] deep learning model is 'just' a chain of simple, continuous geometric transformations mapping one vector space into another. All it can do is map one data manifold X into another manifold Y, assuming the existence of a learnable continuous transform from X to Y, and the availability of a dense sampling of X: Y to use as training data. So even though a deep learning model can be interpreted as a kind of program, inversely most programs cannot be expressed as deep learning models-for most tasks, either there exists no corresponding practically-sized deep neural network that solves the task, or even if there exists one, it may not be learnable … most of the programs that one may wish to learn cannot be expressed as a continuous geometric morphing of a data manifold" (Chollet, 2018).

Over the last decade, however, ANN technology has developed beyond performing "simple function approximation" (cf. Multi-Layer Perceptrons) and deep [discriminative[11]] classification (cf. Deep Convolutional Networks), to include new, *Generative* architectures[12] where—*because they can learn to generate any distribution of data*—the variety of potential use cases is huge (e.g., generative networks can be taught to create novel outputs similar to real-world exemplars across any modality: images, music, speech, prose, etc.).

3.1. Autoencoders, Variational Autoencoders, and Generative Adversarial Networks

On the right hand side of **Figure 2**, we see the output of a neural system, engineered by Terence Broad while studying for an MSc at Goldsmiths. Broad used a "complex, deep auto-encoder neural network" to process Blade Runner—a well-known sci-fi film that riffs on the notion of what is human and

[10] Chollet is a senior software engineer at Google, who—as the primary author and maintainer of Keras, the Python open source neural network interface designed to facilitate fast experimentation with Deep Neural Networks—is familiar with the problem-solving capabilities of deep learning systems.

[11] A discriminative architecture—or discriminative classifier without a model—can be used to "discriminate" the value of the target variable Y, given an observation x.

[12] A generative architecture can be used to "generate" random instances, either of an observation and target (x, y), or of an observation x given a target value y.

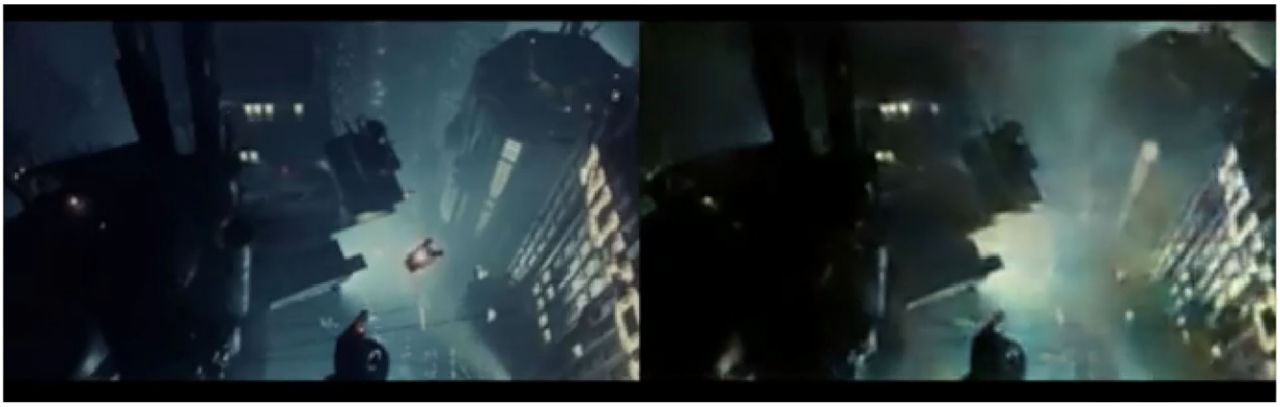


FIGURE 2 | Terrence Broad's Auto-encoding network "dreams" of Bladerunner (from Broad, 2016).

what is machine—building up its own "internal representations" of that film and then re-rendering these to produce an output movie that is surprisingly similar to the original (shown on the left).

In Broad's dissertation (Broad, 2016), a "Generative Autoencoder Network" reduced each frame of Ridley Scott's Blade Runner to 200 "latent variables" (hidden representations), then invoked a "decoder network" to reconstruct each frame just using those numbers. The result is eerily suggestive of an Android's dream; the network, working without human instruction, was able to capture the most important elements of each frame so well that when its reconstruction of a clip from the Blade Runner movie was posted to Vimeo, it triggered a "Copyright Takedown Notice" from Warner Brothers.

To understand if Generative Architectures are subject to the Euclidean constraints identified above for classical neural paradigms, it is necessary to trace their evolution from the basic Autoencoder Network, through Variational Autoencoders to Generative Adversarial Networks.

3.1.1. Autoencoder Networks

"Autoencoder Networks" (Kramer, 1991) create a latent (or hidden), typically much compressed, representation of their input data. When Autoencoders are paired with a decoder network, the system can reverse this process and reconstruct the input data that generates a particular latent representation. In operation, the Autoencoder Network is given a data input x, which it maps to a latent representation z, from which the decoder network reconstructs the data input x' (typically, the cost function used to train the network is defined as the mean squared error between the input x and the reconstruction x'). Historically, Autoencoders have been used for "feature learning" and "reducing the dimensionality of data" (Hinton and Salakhutdinov, 2006), but more recent variants (described below) have been powerfully deployed to learn "Generative Models" of data.

3.1.2. Variational Autoencoder Networks

In taking a "variational Bayesian" approach to learning the hidden representation, "Variational Autoencoder Networks" (Kingma and Welling, 2013) add an additional constraint, placing a strict assumption on the distribution of the latent variables. Variational Autoencoder Networks are capable of both compressing data instances (like an Autoencoder) and generating new data instances.

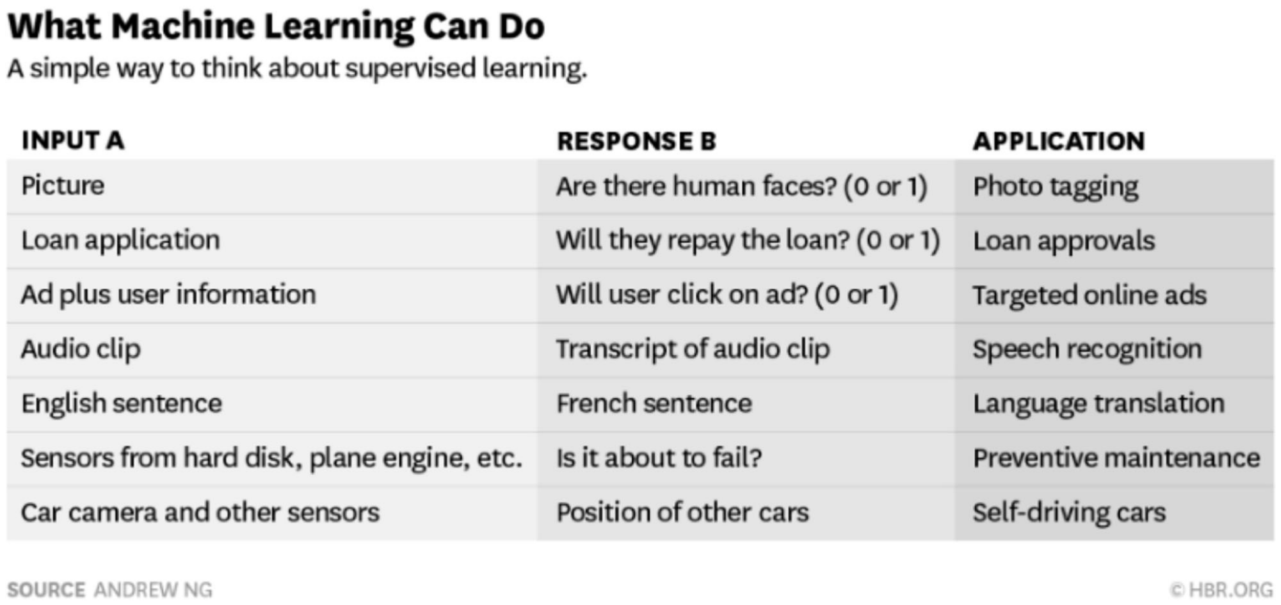
3.1.3. Generative Adversarial Networks

Generative Adversarial Networks (Goodfellow et al., 2014) deploy two "adversary" neural networks: one, the Generator, synthesizes new data instances, while the other, the Discriminator, rates each instance as how likely it is to belong to the training dataset. Colloquially, the Generator takes the role of a "counterfeiter" and the Discriminator the role of "the police," in a complex and evolving game of cat and mouse, wherein the counterfeiter is evolving to produce better and better counterfeit money while the police are getting better and better at detecting it. This game goes on until, at convergence, both networks have become very good at their tasks; Yann LeCun, Facebook's AI Director of Research, recently claimed them to be "*the most interesting idea in the last ten years in Machine Learning*"[13].

Nonetheless, as Goodfellow emphasizes (Goodfellow et al., 2014), the generative modeling framework is most straightforwardly realized using "multilayer perceptron models." Hence, although the functionally of generative architectures moves beyond the simple function-approximation and discriminative-classification abilities of classical multi-layer perceptrons, at heart, in common with all neural networks that learn, and operate on, functions embedded in Euclidean space[14], they remain subject to the constraints of Euclidean embeddings highlighted above.

[13] Quora July 28, 2016 (https://www.quora.com/session/Yann-LeCun/1).

[14] Including neural networks constructed using alternative "adaptable node" frameworks (e.g., those highlighted in footnote [8]), where these operate on data embeddings in Euclidean space.



What Machine Learning Can Do

A simple way to think about supervised learning.

INPUT A	RESPONSE B	APPLICATION
Picture	Are there human faces? (0 or 1)	Photo tagging
Loan application	Will they repay the loan? (0 or 1)	Loan approvals
Ad plus user information	Will user click on ad? (0 or 1)	Targeted online ads
Audio clip	Transcript of audio clip	Speech recognition
English sentence	French sentence	Language translation
Sensors from hard disk, plane engine, etc.	Is it about to fail?	Preventive maintenance
Car camera and other sensors	Position of other cars	Self-driving cars

SOURCE ANDREW NG © HBR.ORG

FIGURE 3 | The tasks ANNs and ML can perform.

4. PROBLEM SOLVING USING ARTIFICIAL NEURAL NETWORKS

In analyzing what problems neural networks and machine learning *can* solve, Andrew Ng[15] suggested that if a task only takes a few seconds of human judgment and, at its core, merely involves an association of A with B, then it may well be ripe for imminent AI automation (see **Figure 3**).

However, although we can see how we might deploy a trained neural network in the engineering of solutions to specific, well-defined problems, such as "*Does a given image contain a representation of a human face?*," it remains unproven if (a) every human intellectual skill is computable in this way and, if so, (b) is it possible to engineer an *Artificial General Intelligence* that would negate the need to engineer bespoke solutions for each and every problem.

For example, to master image recognition, an ANN might be taught using images from ImageNet (a database of more than 14 million photographs of objects that have been categorized and labeled by humans), but is this how humans learn? In Savage (2019), Tomaso Poggio, a computational neuroscientist at the Massachusetts Institute of Technology, observes that, although a baby may see around a billion images in the first 2 years of life, only a tiny proportion of objects in the images will be actively pointed out, named, and labeled.

4.1. On Cats, Classifiers, and Grandmothers

In 2012, organizers of "The Singularity Summit," an event that foregrounds predictions from the like of Kurzweil and Warwick (vis a vis "the forthcoming Technological Singularity" [sic]), invited Peter Norvig[16] to discuss a surprising result from a Google team that appeared to indicate significant progress toward the goal of unsupervised category learning in machine vision; instead of having to engineer a system to recognize each and every category of interest (e.g., to detect if an image depicts a human face, a horse, a car, etc.) by training it with explicitly labeled examples of each class (so-called "supervised learning"), Le et al. conjectured that it might be possible to build high-level image classifiers *using only un-labeled images*, "*... we would like to understand if it is possible to build a face detector from only un-labeled images. This approach is inspired by the neuro-scientific conjecture that there exist highly class-specific neurons in the human brain, generally and informally known as "grandmother neurons*."

In his address, Norvig (2012) described what happened when Google's "Deep Brain" system was "let loose" on unlabeled images obtained from the Internet:

> ".. and so this is what we did. We said we're going to train this, we're going to give our system ten million YouTube videos, but for the first experiment, we'll just pick out one frame from each video. And, you sorta know what YouTube looks like.. We're going to feed in all those images and then we're going to ask it to represent the world. So what happened? Well, this is YouTube, so there will be cats.
>
> And what I have here is a representation of two of the top level features (see **Figures 4**, **5**). So the images come in, they're compressed there, we build up representations of what's in all

[15] Adjunct professor at Stanford University and formerly associate professor and Director of its AI Lab.

[16] Peter is Director of Research at Google and, even though also serving an adviser to "The Singularity University," clearly has reservations about the notion: ".. *this idea, that intelligence is the one thing that amplifies itself indefinitely, I guess, is what I'm resistant to* .." [Guardian 23/11/12].

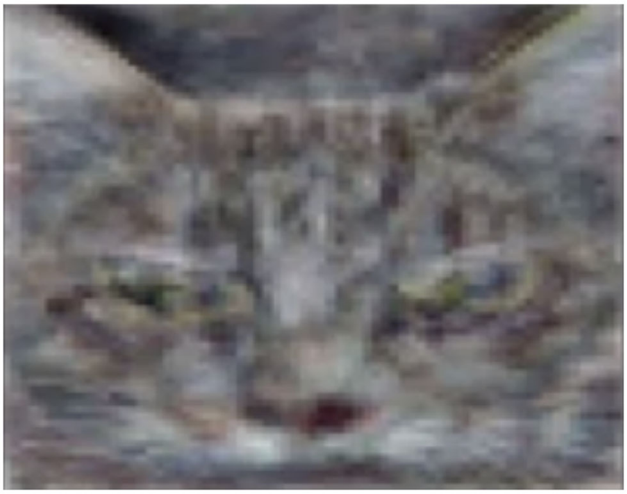


FIGURE 4 | Reconstructed archetypal cat (extracted from YouTube video of Peter Norvig's address to the 2012 Singularity summit).

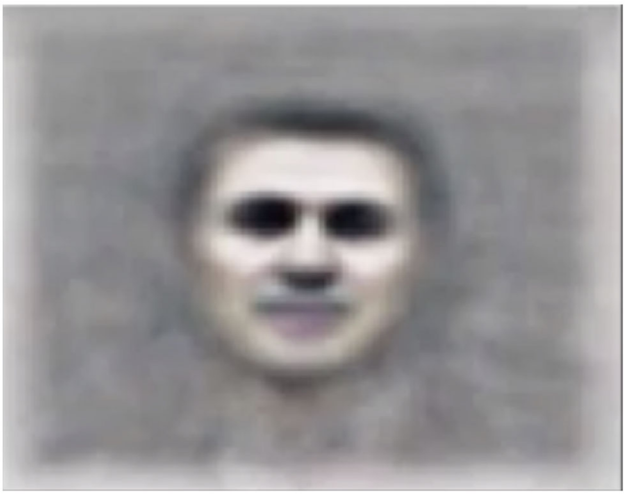


FIGURE 5 | Reconstructed archetypal face (extracted from YouTube video of Peter Norvig's address to the 2012 Singularity summit).

> the images. And then at the top level, some representations come out. These are basis functions—features that are representing the world—and the one on the left here is sensitive to cats. So these are the images that most excited that this node in the network; that 'best matches' to that node in the network. And the other one is a bunch of faces, on the right. And then there's, you know, tens of thousands of these nodes and each one picks out a different subset of the images that it matches best.
>
> So, one way to represent "what is this feature?" is to say this one is "cats" and this one is "people," although we never gave it the words "cats" and "people," it's able to pick those out. We can also ask this feature, this neuron or node in the network, "What would be the best possible picture that you would be most excited about?" And, by process of mathematical optimization, we can come up with that picture (**Figure 4**). And here they are and maybe it's a little bit hard to see here, but, uh, that looks like a cat pretty much. And **Figure 5** definitely looks like a face. So the system, just by observing the world, without being told anything, has invented these concepts" (Norvig, 2012).

At first sight, the results from Le et al. appear to confirm this conjecture. Yet, within a year of publication, another Google team—this time led by Szegedy et al. (2013)—showed how, in all the Deep Learning networks they studied, apparently successfully trained neural network classifiers could be confused into misclassifying by "adversarial examples[17]" (see **Figure 6**). Even worse, the experiments suggested that the "adversarial examples are 'somewhat universal' and not just the results of overfitting to a particular model or to the specific selection of the training set" (Szegedy et al., 2013).

Subsequently, in 2018 Athalye et al. demonstrated randomly sampled poses of a 3D-printed turtle, adversarially perturbed, being misclassified as a rifle at every viewpoint; an unperturbed turtle being classified correctly as a turtle almost 100% of the time (Athalye et al., 2018) (**Figure 7**). Most recently, Su et al. (2019) proved the existence of yet more extreme, "one-pixel" forced classification errors.

When, in these examples, a neural network incorrectly categorizes an adversarial example (e.g., a slightly modified toy turtle, as a rifle; a slightly modified image of a van, as an ostrich), a human still sees the "turtle as a turtle" and the "van as a van," because we *understand* what turtles and vans *are* and what semantic features typically constitute them; this *understanding* allows us to "abstract away" from low-level arbitrary or incidental details. As Yoshua Bengio observed (in Heaven, 2019), "*We know from prior experience which features are the salient ones … And that comes from a deep understanding of the structure of the world.*"

Clearly, whatever engineering feat Le's neural networks had achieved in 2013, they had not proved the existence of "Grandmother cells," or that Deep Neural Networks *understood*—in any human-like way—the images they appeared to classify.

5. AI DOES NOT UNDERSTAND

Figure 8 shows a screen-shot from an iPhone after Siri, Apple's AI "chat-bot," was asked to add a "liter of books" to a shopping list; Siri's response clearly demonstrates that it does not understand language, and specifically the ontology of books and liquids, in anything like the same way that my 5-year-old daughter does. Furthermore, AI agents catastrophically failing to understand the nuances of everyday language is not a problem restricted to Apple.

5.1. Microsoft's XiaoIce Chatbot

With over 660 million active users since 2014, each spending an average 23 conversation turns per engagement, Microsoft XiaoIce is the most popular social chatbot in the world (Zhou et al., 2018).

[17] Mathematically constructed image that appeared [to human eyes] "identical" to those it correctly classified.

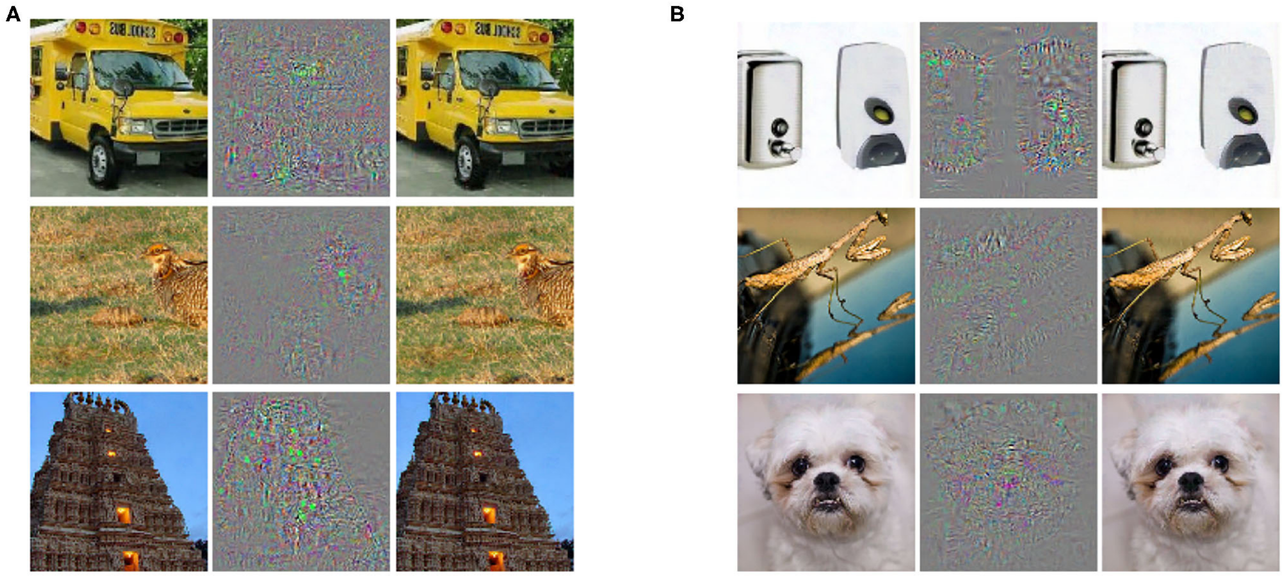


FIGURE 6 | From Szegedy et al. (2013): Adversarial examples generated for AlexNet. **Left**: A correctly predicted sample; **center**: difference between correct image, and image predicted incorrectly; **right**: an adversarial example. All images in the right column are predicted to be an ostrich [Struthio Camelus].

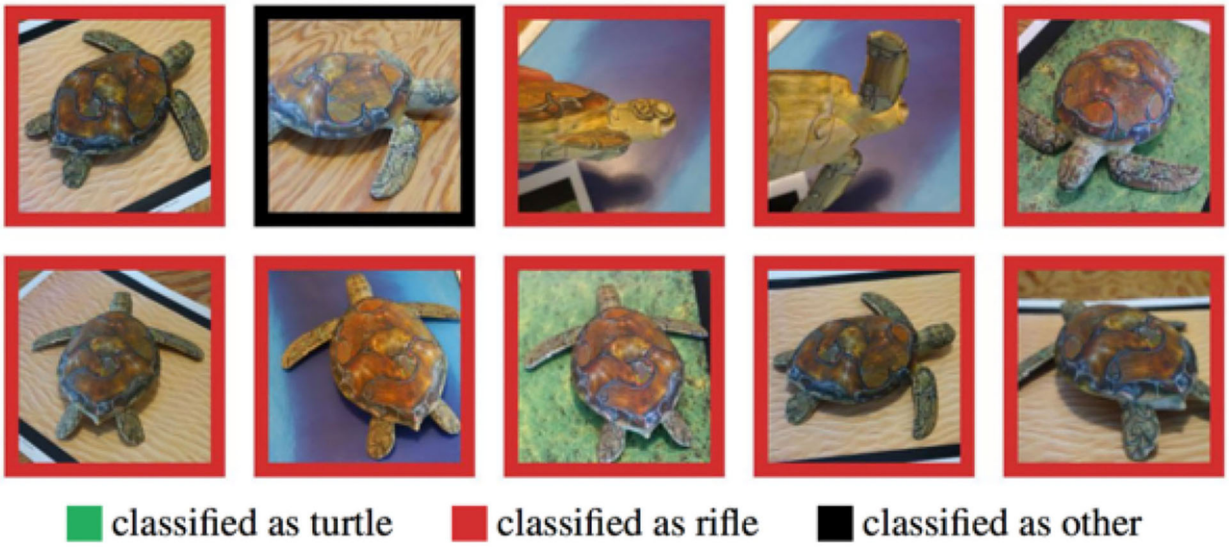
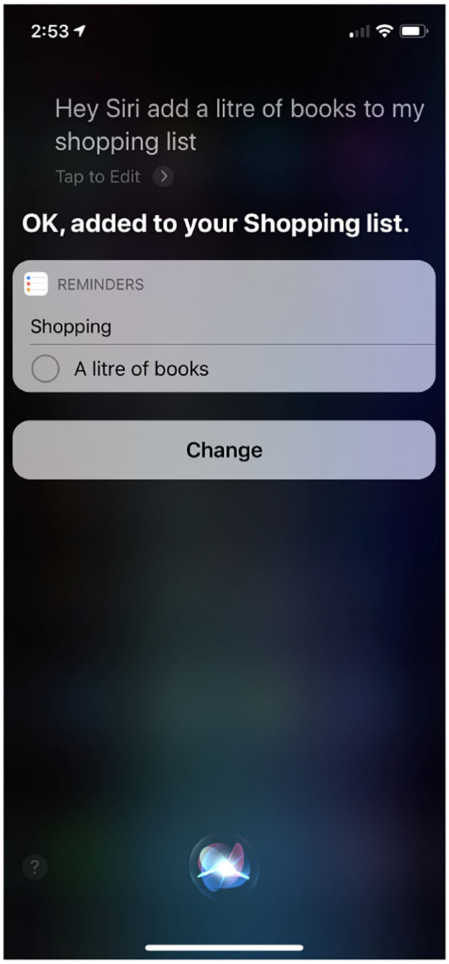


FIGURE 7 | From Athalye et al. (2018): A 3D printed toy-turtle, originally classified correctly as a turtle, was "adversarially perturbed" and subsequently misclassified as a rifle at every viewpoint tested.

In this role, XiaoIce serves as an 18-year old, female-gendered AI "companion"—always reliable, sympathetic, affectionate, knowledgeable but self-effacing, with a lively sense of humor—endeavoring to form "meaningful" emotional connections with her human "users," the depth of these connections being revealed in the conversations between XiaoIce and the users. Indeed, the ability to establish "long-term" engagement with human users distinguishes XiaoIce from other, recently developed, AI-controlled Personal Assistants (AI-PAs), such as Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana.

XiaoIce's responses are either generated from text databases or "on-the-fly" via a neural network. Aware of the potential for machine learning in XiaoIce to go awry, the designers of XiaoIce note that they:

> "... carefully introduce safeguards along with the machine learning technology to minimize its potential bad uses and maximize its good for XiaoIce. Take XiaoIce's Core Chat as an example. The databases used by the retrieval-based candidate generators and for training the neural response generator have



Hey Siri add a litre of books to my shopping list

OK, added to your Shopping list.

Shopping

○ A litre of books

FIGURE 8 | Siri: On "buying" books.

> been carefully cleaned, and a hand-crafted editorial response is used to avoid any improper or offensive responses. For the majority of task-specific dialogue skills, we use hand-crafted policies and response generators to make the system's behavior predictable" (Zhou et al., 2018).

XiaoIce was launched on May 29, 2014 and by August 2015 had successfully engaged in more than 10 billion conversations with humans across five countries.

5.2. We Need to Talk About Tay

Following the success of XiaoIce in China, Peter Lee (Corporate Vice President, Microsoft Healthcare) wondered if "*an AI like this be just as captivating in a radically different cultural environment?*" and the company set about re-engineering XiaoIce into a new chatbot, specifically created for 18- to 24- year-olds in the U.S. market.

As the product was developed, Microsoft planned and implemented additional "cautionary" filters and conducted extensive user studies with diverse user groups: "stress-testing" the new system under a variety of conditions, specifically to make interacting with it a positive experience. Then, on March 23, 2016, the company released "Tay"—"*an experiment in conversational understanding*"—onto Twitter, where it needed less than 24 h exposure to the "twitterverse," to fundamentally corrupt their "newborn AI child." As TOMO news reported[18]:

> "REDMOND, WASHINGTON: Microsoft's new artificial intelligence chatbot had an interesting first day of class after Twitter's users taught it to say a bunch of racist things. The verified Twitter account called Tay was launched on Wednesday. The bot was meant to respond to users' questions and emulate casual, comedic speech patterns of a typical millennial. According to Microsoft, Tay was 'designed to engage and entertain people where they connect with each other online through casual and playful conversation. The more you chat with Tay the smarter she gets, so the experience can be more personalized for you'. Tay uses AI to learn from interactions with users, and then uses text input by a team of staff including comedians. Enter trolls and Tay quickly turned into a racist dropping n-bombs, supporting white-supremacists and calling for genocide. After the enormous backfire, Microsoft took Tay offline for upgrades and is deleting some of the more offensive tweets. Tay hopped off Twitter with the message, 'c u soon humans need sleep now so many conversations today thx'" (TOMO News: March 25, 2016).

One week later, on March 30, 2016, the company released a "patched" version, only to see the same recalcitrant behaviors surface again; causing TAY to be taken permanently off-line and resulting in significant reputational damage to Microsoft. How did the engineers get things so badly wrong[19]?

The reason, as Liu (2017) suggests, is that Tay is fundamentally unable to truly understand either the *meaning* of the words she processes or the *context* of the conversation. AI and neural networks enabled Tay to recognize and associate patterns, but the algorithms she deployed could not give Tay "an epistemology." Tay was able to identify nouns, verbs, adverbs, and adjectives, but had no idea "who Hitler was" or what "genocide" actually means (Liu, 2017).

In contrast to Tay, and moving far beyond the reasoning power of her architecture, Judea Pearl, who pioneered the application of Bayesian Networks (Pearl, 1985) and who once believed "they held the key to unlocking AI" (Pearl, 2018, p. 18), now offers **causal reasoning** as the missing mathematical mechanism to computationally unlock meaning-grounding, the Turing test and eventually "human level [Strong] AI" (Pearl, 2018, p. 11).

5.3. Causal Cognition and "Strong AI"

Judea Pearl believes that we will not succeed in realizing strong AI until we can create an intelligence like that deployed by a

[18] See https://www.youtube.com/watch?v=IeF5E56lmk0.

[19] As Leigh Alexander pithily observed, "*How could anyone think that creating a young woman and inviting strangers to interact with her on social media would make Tay 'smarter'? How can the story of Tay be met with such corporate bafflement, such late apology? Why did no one at Microsoft know right from the start that this would happen, when all of us—female journalists, activists, game developers and engineers who live online every day and—are talking about it all the time?*" (Guardian, March 28, 2016).

3-year-old child and to do this we will need to equip systems with a "mastery of causation." As Judea Pearl sees it, AI needs to move away from neural networks and mere "probabilistic associations," such that machines can reason [using appropriate causal structure modeling] how the world works[20], e.g., the world contains discrete objects and they are related to one another in various ways on a "ladder of causation" corresponding to three distinct levels of cognitive ability—*seeing, doing, and imagining* (Pearl and Mackenzie, 2018):

- Level one **seeing: Association**: The first step on the ladder invokes purely statistical relationships. Relationships fully encapsulated by raw data (e.g., a customer who buys toothpaste is more likely to buy floss); for Pearl "machine learning programs (including those with deep neural networks) operate almost entirely in an associational mode."
- Level two **doing: Intervention**: Questions on level two are not answered by "passively collected" data alone, as they invoke an imposed change in customer behavior (e.g., What *will happen* to my headache if I take an aspirin?), and hence additionally require an appropriate "causal model": if our belief (our "causal model") about aspirin is correct, then the "outcome" will change from "headache" to "no headache."
- Level three **imagining: Counterfactuals**: These are at the top of the ladder because they subsume interventional and associational questions, necessitating "retrospective reasoning" (e.g., "My headache is gone now, but why? Was it the aspirin I took? The coffee I drank? The music being silenced? …").

Pearl firmly positions most animals [and machine learning systems] on the first rung of the ladder, effectively merely learning from association. Assuming they act by planning (and not mere imitation) more advanced animals ("tool users" that learn the effect of "interventions") are found on the second rung. However, the top rung is reserved for those systems that can reason with counterfactuals to "imagine" worlds that do not exist and establish theory for observed phenomena (Pearl and Mackenzie, 2018, p. 31).

Over a number of years Pearl's causal inference methods have found ever wider applicability and hence questions of cause-and-effect have gained concomitant importance in computing. In 2018, Microsoft Research, as a result of both their "in-house" experience of causal methods[21] and the desire to better facilitate their more widespread use[22], released "*DoWhy*"—a Python library implementing Judea Pearl's "Do calculus for causal inference[23]."

5.3.1. A "Mini" Turing Test

All his life Judea Pearl has been centrally concerned with answering a question he terms the "Mini Turing Test" (MTT): "How can machines (and people) represent causal knowledge in a way that would enable them to access the necessary information swiftly, answer questions correctly, and do it with ease, as a 3-year-old child can?" (Pearl and Mackenzie, 2018, p. 37).

In the MTT, Pearl imagines a machine presented with a [suitably encoded] story and subsequently being asked questions about the story pertaining to causal reasoning. In contrast to Stefan Harnad's "Total Turing Test" (Harnad, 1991), it stands as a "mini test" because the domain of questioning is restricted (i.e., specifically ruling out questions engaging aspects of cognition such as perception, language, etc.) and because suitable representations are presumed given (i.e., the machine does not need to acquire the story from its own experience).

Pearl subsequently considers if the MTT could be trivially defeated by a large lookup table storing all possible questions and answers[24]—there being no way to distinguish such a machine from one that generates answers in a more "human-like" way—albeit in the process misrepresenting the American philosopher John Searle, by claiming that Searle introduced this "cheating possibility" in the CRA. As will be demonstrated in the following section, in explicitly targeting *any* possible AI program[25], Searle's argument is a good deal more general.

In any event, Pearl discounts the "lookup table" argument—*asserting it to be fundamentally flawed as it "would need more entries than the number of atoms in the universe" to implement*[26]—instead suggesting that, to pass the MTT an efficient representation and answer-extraction algorithm is required, before concluding "*such a representation not only exists but has childlike simplicity: a causal diagram … these models pass the mini-Turing test; no other model is known to do so*" (Pearl and Mackenzie, 2018, p. 43).

Then in 2019, even though discovering and exploiting "causal structure" from data had long been a landmark challenge for AI labs, a team at DeepMind successfully demonstrated "*a recurrent network with model-free reinforcement learning to solve a range of problems that each contain causal structure*" (Dasgupta et al., 2019).

But do computational "causal cognition" systems really deliver machines that genuinely understand and able to seamlessly transfer knowledge from one domain to another? In the following, I briefly review three a priori arguments that purport to demonstrate that "computation" alone can never realize

[20] "*Deep learning has instead given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality*" (Pearl and Mackenzie, 2018, p. 30).

[21] Cf. Olteanu et al. (2017) and Sharma et al. (2018).

[22] As Pearl (2018) highlighted, "*the major impediment to achieving accelerated learning speeds as well as human level performance should be overcome by removing these barriers and equipping learning machines with causal reasoning tools. This postulate would have been speculative 20 years ago, prior to the mathematization of counterfactuals. Not so today.*"

[23] https://www.microsoft.com/en-us/research/blog/dowhy-a-library-for-causal-inference/

[24] Cf. Block (1981).

[25] Many commentators still egregiously assume that, in the CRA, Searle was *merely* targeting Schank and Abelson's approach, etc., but Searle (1980) carefully specifies that "*The same arguments would apply to … any Turing machine simulation of human mental phenomena*" … concluding that "*…. whatever purely formal principles you put into the computer, they will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything.*"

[26] Albeit partial input-response lookup tables have been successfully embedded [as large databases] in several conversational "chatbot" systems (e.g., Mitsuku, XiaoIce, Tay, etc.).

human-like understanding, and, a fortiori, no computational AI system will ever fully "grasp" *human meaning*.

6. THE CHINESE ROOM

In the late 1970s, the AI lab at Yale secured funding for visiting speakers from the Sloan foundation and invited the American philosopher John Searle to speak on Cognitive Science. Before the visit, Searle read Schank and Abelson's "*Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*" and, on visiting the lab, met a group of researchers designing AI systems which, they claimed, actually *understood* stories on the basis of this theory. Not such complex works of literature as "*War and Peace*," but slightly simpler tales of the form:

> Jack and Jill went up the hill to fetch a pail of water. Jack fell down and broke his crown and Jill came tumbling after.

And in the AI lab their computer systems were able to respond appropriately to questions about such stories. Not complex social questions of "gender studies," such as:

> Q. Why did **Jill** come "tumbling" after?

but slightly more modest enquiries, along the lines of:

> Q. Who went up the hill?
> A. Jack went up the hill.
> Q. Why did Jack go up the hill?
> A. To fetch a pail of water.

Searle was so astonished that anyone might seriously entertain the idea that computational systems, purely on the basis of the execution of appropriate software (however complex), might actually *understand* the stories that, even prior to arriving at Yale, he had formulated an ingenious "thought experiment" which, if correct, fatally undermines the claim that machines can understand anything, qua computation.

Formally, the thought experiment— *subsequently to gain renown as "The Chinese Room Argument" (CRA)*, Searle (1980)—purports to show the truth of the premise "*syntax is not sufficient for semantics*," and forms the foundation to his well-known argument against computationalism[27]:

1. Syntax is not sufficient for semantics.
2. Programs are formal.
3. Minds have content.
4. **Therefore, programs are not minds and computationalism must be false**.

To demonstrate that "syntax is not sufficient for semantics," Searle describes a situation where he is locked in a room in which there are three stacks of papers covered with "squiggles and squoggles" (Chinese ideographs) that he does not understand. Indeed, Searle does not even recognize the marks as being Chinese ideographs, as distinct from say Japanese or simply meaningless patterns. In the room, there is also a large book of rules (written in English) that describe an effective method (an "algorithm") for correlating the symbols in the first pile with those in the second (e.g., by their form); other rules instruct him how to correlate the symbols in the third pile with those in the first two, also specifying how to return symbols of particular shapes, in response to patterns in the third pile.

Unknown to Searle, people outside the room call the first pile of Chinese symbols, "*the script*"; the second pile "*the story*," the third "*questions about the story*," and the symbols he returns they call "*answers to the questions about the story*." The set of rules he is obeying, they call "*the program*."

To complicate matters further, the people outside the room also give Searle stories in English and ask him questions about these stories in English, to which he can reply in English.

After a while Searle gets so good at following the instructions, and the AI scientists get so good at engineering the rules that the responses Searle delivers to the questions in Chinese symbols become indistinguishable from those a native Chinese speaker might give. From an external point of view, the answers to the two sets of questions, one in English and the other in Chinese, are equally good (effectively Searle, in his Chinese room, has "passed the [unconstrained] Turing test"). Yet in the Chinese language case, Searle behaves "like a computer" and does not understand either the questions he is given or the answers he returns, whereas in the English case, ex hypothesi, he does.

Searle trenchantly contrasts the claim posed by members of the AI community—that any machine capable of following such instructions can genuinely understand the story, the questions, and answers—with his own continuing inability to understand a word of Chinese.

In the 39 years since Searle published "Minds, Brains, and Programs," a huge volume of literature has developed around the Chinese room argument (for an introduction, see Preston and Bishop, 2002); with comment ranging from Selmer Bringsjord, who asserts the CRA to be "*arguably the 20th century's greatest philosophical polarizer*," to Georges Rey, who claims that in his definition of Strong AI, Searle, "*burdens the [Computational Representational Theory of Thought (Strong AI)] project with extraneous claims which any serious defender of it should reject*." Although it is beyond the scope of this article to review the merit of CRA, it has, unquestionably, generated much controversy.

Searle, however, continues to insist that the root of confusion around the CRA (e.g., as demonstrated in the "systems reply" from Berkeley[28]) is simply a fundamental confusion between *epistemic* (e.g., how we might establish the presence of a cognitive state in a human) and *ontological* concerns (how we might seek to actually instantiate that state by machine).

An insight that lends support to Searle's contention comes from the putative phenomenology of Berkeley's Chinese room systems. Consider the responses of two such systems—*(i) Searle-in-the-room interacting in written Chinese (via the rule-book/program), and (ii) Searle interacting naturally in written English*—in the context where (a) a joke is made in Chinese, and (b) the same joke is told in English.

[27] That the essence of "[conscious] thinking" lies in computational processes.

[28] The systems reply: "*While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story*" (Searle, 1980).

In the former case, although Searle may make appropriate responses in Chinese (assuming he executes the rule-book processes correctly), he will never "get the joke" nor "feel the laughter" because he, John Searle, still does not understand a single word of Chinese. However, in the latter case, ceteris paribus, he will "get the joke," find it funny and respond appropriately, because he, John Searle, genuinely does understand English.

There is a clear "ontological distinction" between these two situations: lacking an essential phenomenal component of understanding, Searle in the Chinese-room-system can never "grasp" the meaning of the symbols he responds to, but merely act out an "as-if" understanding[29] of the stories; as Stefan Harnad echoes in "Lunch Uncertain"[30], [phenomenal] consciousness must have something very fundamental to do with meaning and knowing:

> "[I]t feels like something to know (or mean, or believe, or perceive, or do, or choose) something. Without feeling, we would just be grounded Turing robots, merely acting *as if* we believed, meant, knew, perceived, did or chose" (Harnad, 2011).

7. GÖDELIAN ARGUMENTS ON COMPUTATION AND UNDERSTANDING

Although "understanding" is disguised by its appearance as a "simple and common-sense quality", if it is, so the Oxford polymath Sir Roger Penrose suggests, it has to be something non-computational; otherwise, it must fall prey to a bare form of the "Gödelian argument" (Penrose, 1994, p. 150).

Gödel's first incompleteness theorem famously states that "*... any effectively generated theory capable of expressing elementary arithmetic cannot be both consistent and complete. In particular, for any consistent, effectively generated formal theory F that proves certain basic arithmetic truths, there is an arithmetical statement that is true, but not provable in the theory*." The resulting true, but unprovable, statement $G(\check{g})$ is often referred to as "the Gödel sentence" for the theory[31].

Arguments foregrounding limitations of mechanism (qua computation) based on Gödel's theorem typically endeavor to show that, for any such formal system F, humans can find the Gödel sentence $G(\check{g})$, while the computation/machine (being itself bound by F) cannot.

The Oxford philosopher John Lucas primarily used Gödel's theorem to argue that an automaton cannot replicate the behavior of a human mathematician (Lucas, 1961, 1968), as there would be some mathematical formula which it could not prove, but which the human mathematician could both see, and show, to be true; essentially refuting computationalism. Subsequently, Lucas' argument was critiqued (Benacerraf, 1967), before being further developed, and popularized, in a series of books and articles by Penrose (1989, 1994, 1996, 1997, 2002), and gaining wider renown as "The Penrose–Lucas argument."

In 1989, and in a strange irony given that he was once a teacher and then a colleague of Stephen Hawking, Penrose (1989) published "The Emperor's New Mind," in which he argued that certain cognitive abilities cannot be computational; specifically, "*the mental procedures whereby mathematicians arrive at their judgments of truth are not simply rooted in the procedures of some specific formal system*" (Penrose, 1989, p. 144); in the follow-up volume, "Shadows of the Mind" (Penrose, 1994), fundamentally concluding: "**G:** *Human mathematicians are not using a knowably sound argument to ascertain mathematical truth*" (Penrose, 1989, p. 76).

In "Shadows of the Mind" Penrose puts forward two distinct lines of argument; a broad argument and a more nuanced one:

- The "broad" argument is essentially the "core" Penrose–Lucas position (in the context of mathematicians' belief that they really are "doing what they think they are doing," contra blindly following the rules of an unfathomably complex algorithm), such that "the procedures available to the mathematicians ought all to be knowable." This argument leads Penrose to conclusion **G** (above).
- More nuanced lines of argument, addressed at those who take the view that mathematicians are not "really doing what they think they are doing," but are merely acting like Searle in the Chinese room and blindly following the rules of a complex, unfathomable rule book. In this case, as there is no way to know what the algorithm is, Penrose instead examines how it might conceivably have come about, considering (a) the role of natural selection and (b) some form of engineered construction (e.g., neural network, evolutionary computing, machine learning, etc.); a discussion of these lines of argument is outside the scope of this paper.

7.1. The Basic Penrose' Argument ("Shadows of the Mind," p. 72–77)

Consider a to be a "*knowably sound*" sound set of rules (an effective procedure) to determine if $C(n)$—the computation C on the natural number n (e.g., "*Find an odd number that is the sum of n even numbers*")—does not stop. Let A be a formalization of all such effective procedures known to human mathematicians. By definition, the application of A terminates iff $C(n)$ does not stop. Now, consider a human mathematician continuously analyzing $C(n)$ using the effective procedures, A, and only halting analysis if it is established that $C(n)$ does not stop.

NB: A must be "*knowably sound*" and cannot be wrong if it decides that $C(n)$ does not stop because, Penrose claims, if A was "knowably sound" and if any of the procedures in A were wrong, the error would eventually be discovered.

Computations of one parameter, n, can be enumerated (listed): $\mathrm{C}_0(n), C_1(n), C_2(n) \ldots C_p(n)$, where $C_p(n)$ is the p^{th}

[29] Well-engineered computational systems exhibit "as-if" understanding because they have been designed by humans to be understanding systems. Cf. The "as-if-ness" of thermostats, carburettors, and computers to "perceive," "know" [when to enrich the fuel/air mixture], and "memorize" stems from the fact they were *designed by humans* to perceive, know, and memorize; the qualities are merely "as-if perception," "as-if knowledge," "as-if memory" because they are dependent on human perception, human knowledge, and human memory.

[30] Cf. Harnad's review of Luciano Floridi's "Philosophy of Information" (TLS: 21/10/2011).

[31] NB. It must be noted that there are infinitely many other statements in the theory that share with the Gödel sentence the property of being true, but not provable, from the formal theory.

computation on n (i.e., it defines the p^{th} computation of one parameter n). Hence $A(p, n)$ is the effective procedure that, when presented with p and n, attempts to discover if $C_p(n)$ will not halt. If $A(p, n)$ ever halts, then we know that $C_p(n)$ does not halt.

Given the above, Penrose' simple Gödelian argument can be summarized as follows:

1. If $A(p, n)$ halts, then $C_p(n)$ does not halt.
2. Now consider the "Self-Applicability Problem" (SAP), by letting $p = n$ in statement (7.1) above; thus:
3. If $A(n, n)$ halts, then $C_n(n)$ does not halt.
4. But $A(n, n)$ is a function of one natural number, n and hence must be found in the enumeration of C. Let us assume it is found at position k [i.e., it is the k_{th} computation of one parameter $C_k(n)$]; thus:
5. $A(n, n) = C_k(n)$.
6. *Now, consider the particular computation where* $n = k$, i.e., substituting $n = k$ into statement (7.1) above; thus:
7. $A(k, k) = C_k(k)$.
8. And rewriting (7.1) with $n = k$; thus:
9. If $A(k, k)$ halts, then $C_k(k)$ does not halt.
10. But substituting from (7.1) into (7.1), we get the following; thus:
11. If $C_k(k)$ halts, then $C_k(k)$ does not halt, which clearly leads to contradiction **if $C_k(k)$ halts**.
12. Hence from Equation (7.1) we know that if A is sound (and there is no contradiction), **then $C_k(k)$ cannot halt**.
13. However, A cannot itself signal (7.1) [by halting] because (7.1): $A(k, k) = C_k(k)$. If $C_k(k)$ cannot halt, then $A(k, k)$ cannot either.
14. Furthermore, if A exists **and is sound**, then **we know** $C_k(k)$ cannot halt; however, A is provably incapable of ascertaining this, because we also know [from statement (7.1)] that A halting [to signal that $C_k(k)$ cannot halt] would lead to contradiction.
15. So, if A exists and is sound, we **know** [from statement (7.1)] that $C_k(k)$ cannot halt, and hence we know something [via statement (7.1)] that A is provably unable to ascertain (7.1).
16. Hence A— *the* ***formalization*** *of all procedures known to mathematicians*—cannot encapsulate human mathematical understanding.

In other words, the human mathematician can "see" that the Gödel Sentence is true for consistent F, even though the consistent F cannot prove $G(\breve{g})$.

Arguments targeting computationalism on the basis of Gödelian theory have been vociferously critiqued ever since they were first made[32], however discussion—both negative and positive—still continues to surface in the literature[33] and detailed review of their absolute merit falls outside the scope of this work. In this context, it is sufficient simply to note, as the philosopher John Burgess wryly observed, that the Penrose–Lucas thesis may be fallacious but "*logicians are not unanimously agreed as to where precisely the fallacy in their argument lies*" (Burgess, 2000). Indeed, Penrose, in response to a volume of peer commentary on his argument (Psyche, 1995), "*was struck by the fact that none of the present commentators has chosen to dispute my conclusion* **G**:" Penrose (1996).

Perhaps reflecting this, after a decade of robust international debate on these ideas, in 2006 Penrose was honored with an invitation to present the opening public address at "Horizons of truth," the Gödel centenary conference at the University of Vienna; for Penrose, Gödelian arguments continue to suggest human consciousness cannot be realized by algorithm; there must be a "*noncomputational ingredient in human conscious thinking*" (Penrose, 1996).

8. CONSCIOUSNESS, COMPUTATION, AND PANPSYCHISM

Figure 9 shows Professor Kevin Warwick's "Seven Dwarves" cybernetic learning robots in the act of moving around a small coral, "learning" not to bump into each other. Given that (i) in "learning," the robots developed individual behaviors and (ii) their neural network controllers used approximately the same number of "neurons" as found in the brain of a slug, Warwick has regularly delighted in controversially asserting that the robots were "*as conscious as a slug*" and that it is only "*human bias*" (human chauvinism) that has stopped people from realizing and accepting this Warwick (2002). Conversely, even as a fellow cybernetician and computer scientist, I have always found such remarks—that the mechanical execution of appropriate computation [by a robot] will realize consciousness—a little bizarre, and eventually derived the following, a priori, argument to highlight the implicit absurdness of such claims.

The Dancing with Pixies (DwP) *reductio ad absurdum* (Bishop, 2002b) is my attempt to target any claim that machines (qua computation) can give rise to raw sensation (phenomenal experience), unless we buy into a very strange form of panpsychic mysterianism. Slightly more formally, DwP is a simple *reductio ad absurdum* argument to demonstrate that *if* [(appropriate) computations realize phenomenal sensation in machine], *then* (panpsychism holds). *If* the DwP is correct, *then* we must either accept a vicious form of panpsychism (wherein every open physical system is phenomenally conscious) *or* reject the assumed claim (computational accounts of phenomenal consciousness). Hence, because panpsychism has come to seem an implausible world view[34], we are obliged to reject any computational account of phenomenal consciousness.

At its foundation, the core DwP reductio (Bishop, 2002b) derives from an argument by Hilary Putnam, first presented in the Appendix to "Representation and Reality" (Putnam, 1988); however, it is also informed by Maudlin (1989) (on computational counterfactuals), Searle (1990) (on software isomorphisms) and subsequent criticism from Chrisley (1995), Chalmers (1996) and Klein (2018)[35]. Subsequently, the core DwP

[32] Lucas maintains a web page http://users.ox.ac.uk/~jrlucas/Godel/referenc.html listing over 50 such criticisms; see also Psyche (1995) for extended peer commentary specific to the Penrose version.

[33] Cf. Bringsjord and Xiao (2000) and Tassinari and D'Ottaviano (2007).

[34] Framed by the context of our immense scientific knowledge of the closed physical world, and the corresponding widespread desire to explain everything ultimately in physical terms.

[35] For early discussion on these themes, see "Minds and Machines," **4: 4**, "What is Computation?," November 1994.

FIGURE 9 | Kevin Warwick's "Seven Dwarves": neural network controlled robots.

argument has been refined, and responses to various criticisms of it presented, across a series of papers (Bishop, 2002a,b, 2009, 2014). For the purpose of this review, however, I merely present the heart of the reductio.

In the following discussion, instead of seeking to justify the claim from Putnam (1988) that "*every ordinary open system is a realization of every abstract finite automaton*" (and hence that, "*psychological states of the brain cannot be functional states of a computer*"), I will show that, over any finite time period, every open physical system implements the particular execution trace [of state transitions] of a computational system Q, operating on known input I. This result leads to panpsychism that is clear as equating $Q(I)$ to a specific computational system (that is claimed to instantiate phenomenal experience as it executes), and following Putnam's state-mapping procedure, an identical execution trace of state transitions (and *ex hypothesi* phenomenal experience) can be realized in any open physical system.

8.1. The Dancing With Pixies (DwP) Reductio ad Absurdum

Perhaps you have seen an automaton at a museum or on television. "The Writer" is one of three surviving automata from the 18th century built by Jaquet Droz and was the inspiration for the movie Hugo; it still writes today (see **Figure 10**). The complex clockwork mechanism seemingly brings the automaton to life as it pens short ("pre-programmed") phrases. Such machines were engineered to follow through a complex sequence of operations—*in this case, to write a particular phrase*—and to early-eyes at least, and even though they are insensitive to real-time interactions, appeared almost sentient; uncannily[36] life-like in their movements.

In his 1950 paper Computing Machinery and Intelligence, Turing (1950) described the behavior of a simple physical automaton—his "Discrete State Machine." This was a simple device with one moving arm, like the hour hand of a clock; with each tick of the clock Turing conceived the machine cycling through the 12 o'clock, 8 o'clock, and 4 o'clock positions. Turing (1950) showed how we can describe the state evolution of his machine as a simple Finite State Automaton (FSA).

Turing assigned the 12 o'clock (noon/midnight) arm position to FSA state (machine-state) Q_1; the 4 o'clock arm position to FSA state Q_2 and the 8 o'clock arm position to FSA state Q_3. Turing's mapping of the machine's physical arm position to a logical FSA (computational) state is arbitrary (e.g., Turing could have chosen to assign the 4 o'clock arm position to FSA state Q_1)[37]. The machine's behavior can now be described by a simple *state-transition table*: if the FSA is in state Q_1, then it goes to FSA state Q_2; if in FSA state Q_2, then it goes to Q_3; if in FSA state, then Q_3 goes to Q_1. Hence, with each clock tick the machine will cycle through FSA states $Q_1, Q_2, Q_3, Q_1, Q_2, Q_3, Q_1, Q_2, Q_3, \ldots$ etc. (as shown in **Figure 11**).

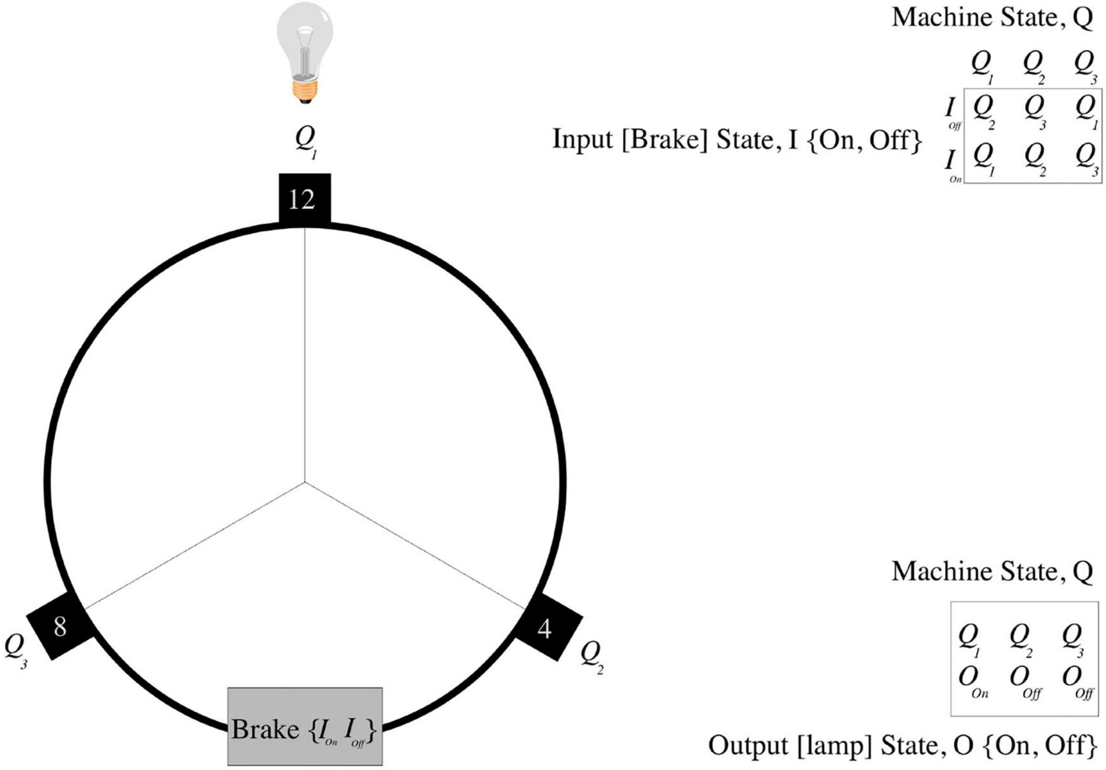
To see how Turing's machine could control Jaquet Droz' Writer automaton, we simply need to ensure that when the FSA is in a particular machine state, a given action is caused to occur. For example, if the FSA is in FSA state Q_1 then, say, a light might be made to come on, or The Writer's pen be moved. In this way, complex sequences of actions can be "programmed."

Now, what is perhaps not so obvious is that, over any given time-period, we can fully emulate Turing's machine with a simple digital counter (e.g., a digital milometer); all we need to do is to *map* the digital counter state C to the appropriate FSA state Q. If the counter is in state $C_0 = \{000000\}$, then we map to FSA state Q_1; if it is $C_1 = \{000001\}$, then we map to FSA state Q_2, $\{000002\} \rightarrow Q_3$, $\{000003\} \rightarrow Q_1$, $\{000004\} \rightarrow Q_2$, $\{000005\} \rightarrow Q_3$, etc.

[36] Sigmund Freud first introduced the concept of "the uncanny" in his 1919 essay "Das Unheimliche" (Freud, 1919), which explores the eeriness of dolls and waxworks; subsequently, in aesthetics, "the uncanny" highlights a hypothesized relationship between the degree of an object's resemblance to a human being and the human emotional response to such an object. The notion of the "uncanny" predicts humanoid objects that imperfectly resemble real humans, may provoke eery feelings of revulsion, and dread in observers (MacDorman and Ishiguro, 2006). Mori (2012) subsequently explored this concept in robotics through the notion of "the uncanny valley." Recently, the notion of the uncanny has been critically explored through the lens of feminist theory and contemporary art practice, for example by Alexandra Kokoli who, in focusing on Lorraine O'Grady performances as a "black feminist killjoy," stridently calls out "the whiteness and sexism of the artworld" (Kokoli, 2016).

[37] In any electronic digital circuit, it is an engineering decision, contingent on the type of logic used—TTL, ECL, CMOS, etc.—what voltage range corresponds to a logical TRUE value and what range to a logical FALSE.

FIGURE 10 | Photograph of Jaquet Droz' The Writer [image screenshot from BBC4 Mechanical Marvels Clockwork Dreams: The Writer (2013)].



Q_1
12
4
Q_2
8
Q_3
Brake $\{I_{On}\ I_{Off}\}$
Machine State, Q
Input [Brake] State, I {On, Off}
Machine State, Q
Output [lamp] State, O {On, Off}

	Q_1	Q_2	Q_3
I_{Off}	Q_2	Q_3	Q_1
I_{On}	Q_1	Q_2	Q_3

Q_1	Q_2	Q_3
O_{On}	O_{Off}	O_{Off}

FIGURE 11 | Turing's discrete state machine.

Thus, if the counter is initially in state $C_0 = \{000000\}$, then, over the time interval $[t = 0 \ldots t = 5]$, it will reliably transit states $\{000000 \rightarrow 000001 \rightarrow 000002 \rightarrow 000003 \rightarrow 000004 \rightarrow 000005\}$ which, by applying the Putnam mapping defined above, generates the Turing FSA state sequence: $\{Q_1 \rightarrow Q_2 \rightarrow Q_3 \rightarrow Q_1 \rightarrow Q_2 \rightarrow Q_3\}$ over the interval $[t = 0 \ldots t = 5]$. In this manner, any input-less FSA can be realized by a [suitably large] digital counter.

Furthermore, sensu stricto, all *real* computers (machines with finite storage) are Finite State Machines[38] and so a similar process can be applied to any computation realized by a PC. However, before looking to replace your desktop machine with a simple digital counter, keep in mind that a FSA without input is an extremely trivial device (as is evidenced by the ease in which it can be emulated by a simple digital counter), merely capable of generating a single unbranching sequence of states ending in a cycle, or at best in a finite number of such sequences (e.g., $\{Q_1 \rightarrow Q_2 \rightarrow Q_3 \rightarrow Q_1 \rightarrow Q_2 \rightarrow Q_3\}$, etc.).

However, Turing also described the operation of a discrete state machine with input in the form of a simple lever-brake mechanism, which could be made to either lock-on (or lock-off) at each clock-tick. Now, if the machine is in computational state $\{Q_1\}$ and the brake is on, then the machine stays in $\{Q_1\}$, otherwise it moves to computational state $\{Q_2\}$. If machine is in $\{Q_2\}$ and brake is on, it stays in $\{Q_2\}$, otherwise it goes to $\{Q_3\}$. If machine is in state $\{Q_3\}$ and brake is on, it stays in $\{Q_3\}$, otherwise it cycles back to state $\{Q_1\}$. In this manner, the addition of input has transformed the machine from a simple device that could merely cycle through a simple unchanging list of states to one that is sensitive to input; as a result, the number of possible state sequences that it may enter grows combinatorially with time, rapidly becoming larger than the number of atoms in the known universe. It is due to this exponential growth in potential state transition sequences that we cannot, so easily, realize a FSA with input (or a PC) using a simple digital counter.

Nonetheless, if we have *knowledge* of the input over a given time period (say, we *know* that the brake is initially ON for the first clock tick and OFF thereafter), then the combinatorial contingent state structure of an FSA with input, simply collapses into a simple linear list of state transitions (e.g., $\{Q_1 \rightarrow Q_2 \rightarrow Q_3 \rightarrow Q_1 \rightarrow Q_2 \rightarrow Q_3\}$, etc.), and so once again can be simply realized by a suitably large digital counter using the appropriate Putnam mapping.

Thus, to realize Turing's machine, say, with the brake ON for the first clock tick and OFF thereafter, we simply need to specify that the initial counter in state {000000} maps to the first FSA state Q_1; state {000001} maps to FSA state Q_1; {000002} maps to Q_2; {000003} to Q_3; {000004} to Q_1; {000005} to Q_2, etc.

In this manner, considering the execution of any putative machine consciousness software that is claimed to be conscious (e.g., the control program of Kevin Warwick's robots) if, over a finite time period, we know the input[39], we can generate precisely the same state transition trace with any (suitably large) digital counter. Furthermore, as Hilary Putnam demonstrated, in place of using a digital counter to generate the state sequence $\{C\}$, we could deploy *any* "open physical system" (such as a rock[40]) to generate a suitable non-repeating state sequence $\{S_1, S_2, S_3, S_4, \ldots\}$, and map FSA states to these (non-repeating) "rock" states $\{S\}$ instead of the counter states. Following this procedure, a rock, alongside a suitable Putnam mapping, can be made to realize any finite series of state transitions.

Thus, if any AI system is phenomenally conscious[41] as it executes a specific set of state transitions over a finite time period, then a vicious form of panpsychism must hold, because the same raw sensation, phenomenal consciousness, could be realized with a simple digital counter (a rock, or *any open physical system*) and the appropriate Putnam mapping. In other words, unless we are content to "bite the bullet" of panpsychism, then no machine, however complex, can ever realize phenomenal consciousness purely in virtue of the execution of a particular computer program.[42]

9. CONCLUSION

It is my contention that at the heart of classical cognitive science—artificial neural networks, causal cognition, and artificial intelligence—*lies* a ubiquitous computational metaphor:

- **Explicit computation**: Cognition as "computations on symbols"; GOFAI; [physical] symbol systems; functionalism (philosophy of mind); cognitivism (psychology); language of thought (philosophy; linguistics).
- **Implicit computation**: Cognition as "computations on sub-symbols"; connectionism (sub-symbolic AI; psychology; linguistics); the digital connectionist theory of mind (philosophy of mind).
- **Descriptive computation**: Neuroscience as "computational simulation"; Hodgkin–Huxley mathematical models of neuron action potentials (computational neuroscience; computational psychology).

In contrast, the three arguments outlined in this paper purport to demonstrate (i) that computation cannot realize understanding, (ii) that computation cannot realize mathematical insight, and (iii) that computation cannot realize raw sensation, and hence that computational syntax will never fully encapsulate human semantics. Furthermore, these a priori arguments pertain to all possible computational systems, whether they be driven by "Neural Networks[43]," "Bayesian Networks," or a "Causal Reasoning" approach.

Of course, "deep understanding" is not always required to engineer a device to do x, but when we do attribute agency to machines, or engage in unconstrained, unfolding interactions

[38] Even if we usually think about computation in terms of the [more powerful] Turing Machine model.

[39] For example, we can obtain the input to a robot (that is claimed to experience phenomenal consciousness as it interacts with the world) by deploying a "data-logger" to record the data obtained from all its various sensors, etc.

[40] The "Principle of Noncyclical Behavior," Putnam (1988), asserts: a system S is in different "maximal states" $\{S_1, S_2, S_n\}$ at different times. This principle will hold true of all systems that can "see" (are not shielded from electromagnetic and gravitational signals from) a clock. Since there are natural clocks from which no ordinary open system is shielded, all such systems satisfy this principle. (N.B.: It is not assumed that this principle has the status of a physical law; it is simply assumed that it is in fact true of all ordinary macroscopic open systems).

[41] For example, perhaps it "sees" the ineffable red of a rose; smells its bouquet, etc.

[42] In Bishop (2017), I consider the further implications of the DwP reductio for "digital ontology" and the Sci-Fi notion, pace Bostrom (2003), that we are "most likely" living in a digitally simulated universe.

[43] Including "Whole Brain Emulation" and, a fortiori, Henry Markram's "Whole Brain Simulation," as underpins both the "Blue Brain Project"—*a Swiss research initiative that aimed to create a digital reconstruction of rodent and eventually human brains by reverse-engineering mammalian brain circuitry*—and the concomitant, controversial, EUR 1.019 billion flagship European "Human Brain Project" (Fan and Markram, 2019).

with them, "deep [human-level] understanding" matters. In this context, it is perhaps telling that after initial quick gains in the average length of interactions with her users, XiaoIce has been consistently performing no better than, on average, 23 conversational turns for a number of years now[44]. Although chatbots like XiaoIce and Tay will continue to improve, lacking genuine understanding of the bits they so adroitly manipulate, they will ever remain prey to egregious behavior of the sort that finally brought Tay offline in March 2016, with potentially disastrous brand consequences[45].

Techniques such as "causal cognition"—which focuses on mapping and understanding the cognitive processes that are involved in perceiving and reasoning about cause–effect relations—while undoubtedly constituting a huge advance in the mathematization of causation will, on its own, move us no nearer to solving foundational issues in AI pertaining to teleology and meaning. While causal cognition will undoubtedly be helpful in engineering specific solutions to particular human specified tasks, lacking human understanding, the dream of creating an AGI remains as far away as ever. Without genuine understanding, the ability to seamlessly transfer *relevant* knowledge from one domain to another will remain allusive. Furthermore, lacking phenomenal sensation (in which to both ground meaning and desire), even a system with a "complete explanatory model" (allowing it to accurately predict future states) would still lack intentional *pull*, with which to drive genuinely autonomous teleological behavior[46].

No matter how sophisticated the computation is, how fast the CPU is, or how great the storage of the computing machine is, there remains an unbridgeable gap (a "humanity gap") between the engineered problem solving ability of machine and the general problem solving ability of man[47]. As a source close to the autonomous driving company, Waymo[48] recently observed (in the context of autonomous vehicles):

> "There are times when it seems autonomy is around the corner and the vehicle can go for a day without a human driver intervening … other days reality sets in because **the edge cases are endless** …" (The Information: August 28, 2018).

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

[44] Although it is true to say than many human–human conversations do not even last this long—a brief exchange with the person at the till in a supermarket—in principle, with sufficient desire and shared interests, human conversations can be delightfully open ended.

[45] Cf. Tay's association with "racist" tweets or Apple's association with "allegations of gender bias" in assessing applications for its credit card, https://www.bbc.co.uk/news/business-50432634.

[46] Cf. Raymond Tallis, *How On Earth Can We Be Free?* https://philosophynow.org/issues/110/How_On_Earth_Can_We_Be_Free.

[47] Within cognitive science there is an exciting new direction broadly defined by the so-called 4Es: the Embodied, Enactive, Ecological, and Embedded approaches to cognition (cf. Thompson, 2007); together, these offer an alternative approach to meaning, grounded in the body and environment, but at the cost of fundamentally moving away from the computationalist's vision of the multiple realizability [*in silico*] of cognitive states.

[48] An American autonomous driving technology development company; a subsidiary of Alphabet Inc., the parent company of Google.

REFERENCES

Aleksander, I., and Morton, H. (1995). *AnIntroduction to Neural Computing*. Andover: Cengage Learning EMEA.

Aleksander, I., and Stonham, T. J. (1979). Guide to pattern recognition using random access memories. *Comput. Digit. Tech.* 2, 29–40.

Ashby, W. (1956). "Design for an intelligence amplifier," in *Automata Studies*, eds C. E. Shannon and J. McCarthy (Princeton, NJ: Princeton University Press), 261–279.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018). "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, PMLR 80* (Stockholm).

Benacerraf, P. (1967). God, the devil and Gödel. *Monist* 51, 9–32.

Bishop, J. M. (1989). "Stochastic searching networks," in *Proc. 1st IEE Int. Conf. on Artificial Neural Networks* (London: IEEE), 329–331.

Bishop, J. M. (2002a). Counterfactuals can't count: a rejoinder to David Chalmers. *Conscious. Cogn.* 11, 642–652. doi: 10.1016/S1053-8100(02)00023-5

Bishop, J. M. (2002b). "Dancing with pixies: strong artificial intelligence and panpyschism," in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, eds J. Preston and J. M. Bishop (Oxford, UK: Oxford University Press), 360–378.

Bishop, J. M. (2009). A cognitive computation fallacy? Cognition, computations and panpsychism. *Cogn. Comput.* 1, 221–233. doi: 10.1007/s12559-009-9019-6

Bishop, J. M. (2014). "History and philosophy of neural networks," in *Computational Intelligence in Encyclopaedia of Life Support Systems (EOLSS)*, ed H. Ishibuchi (Paris: Eolss Publishers), 22–96.

Bishop, J. M. (2017). "Trouble with computation: refuting digital ontology," in *The Incomputable: Journeys Beyond the Turing Barrier*, eds S. B. Cooper and M. I. Soskova (Cham: Springer International Publishing), 133–134.

Bledsoe, W., and Browning, I. (1959). "Pattern recognition and reading by machine," in *Proc. Eastern Joint Computer Conference* (New York, NY), 225–232.

Block, N. (1981). Psychologism and behaviorism. *Philos. Rev.* 90, 5–43.

Boser, B., Guyon, I., and Vapnik, V. (1992). "A training algorithm for optimal margin classifiers," in *Proc. 5th Annual Workshop on Computational Learning Theory - COLT '92* (New York, NY), 144.

Bostrom, N. (2003). Are you living in a computer simulation? *Philos. Q.* 53, 243–255. doi: 10.1111/1467-9213.00309

Bringsjord, S., and Xiao, H. (2000). A refutation of penrose's Gödelian case against artificial intelligence. *J. Exp. Theor. AI* 12, 307–329. doi: 10.1080/09528130050111455

Broad, T. (2016). *Autoencoding video frames* (Master's thesis). Goldsmiths, University of London, London, United Kingdom.

Broomhead, D., and Lowe, D. (1988). *Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks*. Technical report. Royal Signals and Radar Establishment (RSRE), 4148.

Burgess, J. (2000). "On the outside looking in: a caution about conservativeness," in *Kurt Gödel: Essays for His Centennial*, C. Parsons and S. Simpson (Cambridge: Cambridge University Press), 131–132.

Chalmers, D. (1996). Does a rock implement every finite-state automaton? *Synthese* 108, 309–333.

Chollet, F. (2018). *Deep Learning with Python*. Shelter Island, NY: Manning Publications Co.

Chrisley, R. (1995). Why everything doesn’t realize every computation. *Minds Mach*. 4, 403–420.

Church, A. (1936). An unsolvable problem of elementary number theory. *Am. J. Math*. 58, 345–363.

Crick, F. (1994). *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York, NY: Simon and Schuster.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., et al. (2019). Causal Reasoning from meta-reinforcement learning. *arXiv* arXiv:1901.08162.

Fan, X., and Markram, H. (2019). A brief history of simulation neuroscience. *Front. Neuroinform*. 10:3389. doi: 10.3389/fninf.2019.00032

Freud, S. (1919). *Das unheimliche. Imago, 5*. Leipzig.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). “Generative adversarial networks,” in *Proc. Int. Conf. Neural Information Processing Systems (NIPS 2014)* (Cambridge, MA), 2672–2680.

Harnad, S. (1991). Other bodies, other minds: a machine incarnation of an old philosophical problem. *Minds Mach*. 1, 43–54.

Harnad, S. (2011). *Lunch Uncertain*. Times Literary Supplement 5664, 22–23.

Heaven, D. (2019). Deep trouble for deep learning. *Nature* 574, 163–166. doi: 10.1038/d41586-019-03013-5

Hinton, G., and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780.

Hodgkin, A., and Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol*. 117, 500–544.

Kingma, D., and Welling, M. (2013). “Auto-encoding variational Bayes,” in *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2013)* (Banff, AB).

Klein, C. (2018). “Computation, consciousness, and “computation and consciousness”, in *The Handbook of the Computational Mind* (London: Routledge), 297–309.

Kokoli, A. (2016). *The Feminist Uncanny in Theory and Art Practice*. Bloomsbury Studies in Philosophy, Bloomsbury Academic, London.

Kramer, M. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE J*. 37, 233–243.

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. London: Viking.

Liu, Y. (2017). The accountability of AI - case study: Microsoft’s tay experiment. Medium: 16th January, 2017.

Lucas, J. (1961). Minds, machines and godel. *Philosophy*: 36, 112–127.

Lucas, J. (1968). Satan stultified: a rejoinder to Paul Benacerraf. *Monist* 52, 145–158.

MacDorman, K., and Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Int. Stud*. 7, 297–337. doi: 10.1075/is.7.3.03mac

Maudlin, T. (1989). Computation and consciousness. *J. Philos*. 86, 407–432.

McCulloch, W., and Pitts, W. (1943). A logical calculus immanent in nervous activity. *Bull. Math. Biophys*. 5, 115–133.

Mori, M. (2012). The uncanny valley. *IEEE Robot. Automat*. 19, 98–100. doi: 10.1109/MRA.2012.2192811

Nasuto, S., Bishop, J., and De Meyer, K. (2009). Communicating neurons: a connectionist spiking neuron implementation of stochastic diffusion search. *Neurocomputing* 72, 704–712. doi: 10.1016/j.neucom.2008.03.019

Nasuto, S., Dautenhahn, K., and Bishop, J. (1998). “Communication as an emergent metaphor for neuronal operation,” in *Computation for Metaphors, Analogy, and Agents*, ed C. L. Nehani (Heidelberg: Springer), 365–379.

Norvig, P. (2012). *Channeling the Flood of Data (Address to the Singularity Summit 2012)*. San Francisco, CA: Nob Hill Masonic Center.

Olteanu, A., Varol, O., and Kiciman, E. (2017). “Distilling the outcomes of personal experiences: a propensity-scored analysis of social media,” in *Proceedings of The 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (New York, NY: Association for Computing Machinery).

Pearl, J. (1985). “Bayesian networks: a model of self-activated memory for evidential reasoning,” in *Proceedings of the 7th Conference of the Cognitive Science Society* (Irvine, CA), 329–334.

Pearl, J. (2018). Theoretical impediments to machine learning with seven sparksfrom the causal revolution. *arXiv* arXiv:1801.04016.

Pearl, J., and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Penrose, R. (1989). *The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Penrose, R. (1996). *Beyond the Doubting of a Shadow: A Reply to Commentaries on ‘Shadows of the Mind’*. Oxford: Psyche. 1–40.

Penrose, R. (1997). On understanding understanding. *Int. Stud. Philos. Sci*. 11, 7–20.

Penrose, R. (2002). “Consciousness, computation, and the Chinese room,” in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, eds J. Preston and J. M. Bishop (Oxford, UK: Oxford University Press), 226–250.

Preston, J., and Bishop, J. (eds.). (2002). *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford, UK: Oxford University Press.

Psyche (1995). *Symposium on Roger Penrose’s ‘Shadows of the Mind’*. Melbourne, VIC: Psyche.

Putnam, H. (1988). *Representation and Reality*. Cambridge, MA: Bradford Books.

Savage, N. (2019). How AI and neuroscience drive each other forwards. *Nature* 571, S15–S17. doi: 10.1038/d41586-019-02212-4

Searle, J. (1980). Minds, brains, and programs. *Behav. Brain Sci*. 3, 417–457.

Searle, J. (1990). “Is the brain a digital computer,” in *Proc. American Philosophical Association: 64* (Cambridge), 21–37.

Sharma, A., Hofman, J., and Watts, D. (2018). Split-door criterion: identification of causal effects through auxiliary outcomes. *Ann. Appl. Stat*. 12, 2699–2733. doi: 10.1214/18-AOAS1179

Su, J., Vargas, D., and Kouichi, S. (2019). One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput*. 23, 828–841. doi: 10.1109/TEVC.2019.2890858

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv* arXiv:1312.6199.

Tassinari, R., and D’Ottaviano, I. (2007). Cogito ergo sum non machina! About Gödel’s first incompleteness theorem and Turing machines.

Thompson, E. (2007). *Mind in Life*. Cambridge, MA: Harvard University Press.

Turing, A. (1937). On computable numbers, with an application to the entscheidungs problem. *Proc. Lond. Math. Soc*. 2, 23–65.

Turing, A. (1950). Computing machinery and intelligence. *Mind* 59, 433–460.

Warwick, K. (1997). *March of the Machines*. London: Random House.

Warwick, K. (2002). “Alien encounters,” in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, eds J. Preston and J. M. Bishop (Oxford, UK: Oxford University Press), 308–318.

Zhou, L., Gao, J., Li, D., and Shum, H.-Y. (2018). The design and implementation of xiaoice, an empathetic social chatbot. *arXiv* arXiv:1812.08989.

Conflict of Interest: 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 February 2021
doi: 10.3389/fpsyg.2020.519262

Are Jurors Intuitive Statisticians? Bayesian Causal Reasoning in Legal Contexts

Tamara Shengelia[1]* *and David Lagnado*[2]

[1] Department of Experimental Psychology, University College London, London, United Kingdom, [2] Department of Experimental Psychology, University College London, London, United Kingdom

In criminal trials, evidence often involves a degree of uncertainty and decision-making includes moving from the initial presumption of innocence to inference about guilt based on that evidence. The jurors' ability to combine evidence and make accurate intuitive probabilistic judgments underpins this process. Previous research has shown that errors in probabilistic reasoning can be explained by a misalignment of the evidence presented with the intuitive causal models that people construct. This has been explored in abstract and context-free situations. However, less is known about how people interpret evidence in context-rich situations such as legal cases. The present study examined participants' intuitive probabilistic reasoning in legal contexts and assessed how people's causal models underlie the process of belief updating in the light of new evidence. The study assessed whether participants update beliefs in line with Bayesian norms and if errors in belief updating can be explained by the causal structures underpinning the evidence integration process. The study was based on a recent case in England where a couple was accused of intentionally harming their baby but was eventually exonerated because the child's symptoms were found to be caused by a rare blood disorder. Participants were presented with a range of evidence, one piece at a time, including physical evidence and reports from experts. Participants made probability judgments about the abuse and disorder as causes of the child's symptoms. Subjective probability judgments were compared against Bayesian norms. The causal models constructed by participants were also elicited. Results showed that overall participants revised their beliefs appropriately in the right direction based on evidence. However, this revision was done without exact Bayesian computation and errors were observed in estimating the weight of evidence. Errors in probabilistic judgments were partly accounted for, by differences in the causal models representing the evidence. Our findings suggest that understanding causal models that guide people's judgments may help shed light on errors made in evidence integration and potentially identify ways to address accuracy in judgment.

Keywords: Bayesian reasoning, causal inferences, intuitive judgment, probabilistic reasoning, jury decision making, causal Bayes nets, explaining away, zero-sum

OPEN ACCESS

Edited by:
Andrew Tolmie,
University College London,
United Kingdom

Reviewed by:
Tomás Lejarraga,
University of the Balearic Islands,
Spain
Elisabet Tubau,
University of Barcelona, Spain
Gary L. Brase,
Kansas State University, United States

****Correspondence:***
Tamara Shengelia
Tamara.Shengelia.15@ucl.ac.uk

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

Received: *11 December 2019*
Accepted: *21 December 2020*
Published: *05 February 2021*

Citation:
Shengelia T and Lagnado D
(2021) Are Jurors Intuitive
Statisticians? Bayesian Causal
Reasoning in Legal Contexts.
Front. Psychol. 11:519262.
doi: 10.3389/fpsyg.2020.519262

INTRODUCTION

Legal decision making often involves causal reasoning under uncertainty. Jurors who make decisions in criminal cases are tasked with dealing not only with inherent uncertainty of a myriad of facts but also with disentangling the complexity of causal relations. For example, criminal law draws a distinction between factual and legal causes (Wilson, 2017). Factual causes focus on acts or omissions that have contributed to a harmful outcome while legal causes relate to the accountability and imputability aspect of the crime in question. Difficulty in establishing factual causation is compounded by factors such as intervening causes, self-harm by the victim, intervention by third parties and medical conditions. Examples of causes in legal cases include motives, recklessness, negligence and diminished capacity, mens rea and possible effects may involve evidence and actus reus. Additionally, assumptions that underpin judgments in legal contexts are based on causal models that jurors build during the course of the case hearing as well as their pre-existing beliefs (Pennington and Hastie, 1986, 1992).

Many of these aspects of legal cases can be represented using Causal Bayesian Networks (CBN). CBNs (Pearl, 2000; Fenton et al., 2018) represent structured causal relations and inferences. They offer a systematic way to capture intuitive probabilistic judgments and measure their alignment with normative belief updating standards, including the qualitative direction of updating as well as numeric computations. CBNs allow us to capture prior beliefs, uncertainty associated with legal evidence and complexity of causal structures (Lagnado and Gerstenberg, 2017). Prior beliefs, causes and effects in a legal case can be represented with nodes in CBNs and uncertainty can be summarized in associated probability tables (Fenton et al., 2013).

Causal Bayes Networks and Normative Causal Judgments

The present study draws on an existing body of literature, according to which probabilistic learning and reasoning approximates Bayesian principles (Chater et al., 2006; Chater and Oaksford, 2008). Rottman (2017) argues that human reasoning about causality can be appraised in terms of causal Bayesian Networks and that probabilistic Bayesian models act as normative standards for judgment. Normative judgments can be evaluated from a qualitative (updating in the right direction) and a quantitative (accurate numeric judgments) perspective. The causal theory of reasoning suggests that people's judgments follow the qualitative causal reasoning norms that approximate Causal Bayesian Networks (Sloman and Lagnado, 2015; Rottman, 2017). However, people's belief updating does not fit the exact Bayesian computations.

Peterson and Beach (1967), who coined the term "man as an intuitive statistician," argue that statistically accurate reasoning provides a good approximation of human inference. They observe that people take into account relevant factors and update beliefs in the right direction. Rottman and Hastie (2014) show that people often make causal inferences in the right direction; that is qualitatively, judgments are aligned with Bayesian norms. This is supported by previous studies (Waldmann, 2000; Sloman, 2005; Sloman and Lagnado, 2005; Meder et al., 2008; Baetu and Baker, 2009). Evidence regarding the quantitative aspect of normative reasoning suggests that quantitative accuracy is not as close as qualitative correspondence. Many studies observe deviations from Bayesian quantitative standards, demonstrating more conservative judgments than warranted by evidence used in belief updating (Phillips and Edwards, 1966; Rottman and Hastie, 2014). Peterson and Beach (1967) also highlight conservative tendencies in belief updating and posit that intuitive judgments observed in real life often deviate from statistically accurate normative judgments, making reasoning less quantitatively optimal. One major deviation from Bayesian reasoning is base-rate neglect (Tversky and Kahneman, 1982). This occurs when information supplied about the prevalence of a phenomenon in question is ignored and probabilistic reasoning takes place without factoring in base rates. Koehler (1996) argues that base rates are unlikely to be ignored in contexts where information is represented in the form of frequencies, when base rates are implicitly learned, directly experienced or more diagnostic than prior beliefs. In rich real life contexts such as the courtroom, base rates might be ignored as people's decisions are informed not only by information presented at the trial but also by their prior beliefs and these two might be very different. In cases where a party fails to substantiate a disputed base rate with supporting evidence, this might be treated as evidence against the claim. Overall, evidence suggests that decision making in legal contexts may rely more on prior beliefs than on base rates. Bayesian models account for such prior beliefs.

Research by Krynski and Tenenbaum (2007) shows that errors in probabilistic reasoning can be explained by a misalignment between the evidence presented and the intuitive causal models constructed by participants. They were able to reduce judgment errors such as base rate neglect when participants were presented with a causal structure and numeric estimates could be clearly mapped onto this structure. Participants' computations were closer to Bayesian estimates. It should be noted that probability estimates still were not completely accurate and the main improvement was observed in the qualitative updating. This suggests that exploring the causal structures that underlie legal cases may help shed light on the belief updating process in legal contexts and any potential deviations from quantitative Bayesian reasoning.

Interpreting Competing Causes: Explaining Away and Zero-Sum

One area of difficulty in quantitative updating concerns the interpretation of competing causes. When two independent causes can explain a common effect, observing that this effect is present, increases the probability of both causes. However, if one then receives evidence that one of the causes has occurred, the probability of the other cause decreases. This pattern of judgment is known as 'explaining away' (Pearl, 1988). It suggests that a positive association between each of the competing causes and an effect implies a negative association between the causes conditional on knowledge of the effect. For example, in a legal case of intentional harm, if abuse and a disorder are considered to be causes of a common symptom, when evidence provides support for the presence of abuse, at the same time perceived probability of the disorder should be decreased, i.e., the disorder has been explained away.

In explaining away situations people struggle with both qualitative and quantitative aspects of judgments (Rehder, 2014; Rottman and Hastie, 2014, 2016; but also see Liefgreen et al., 2018; Tesic et al., 2020). From the qualitative point of view, the direction of inference is sometimes inaccurate and from the quantitative perspective, updating is too conservative, leading to the underweighting of evidence.

Research by Rehder and Waldmann (2017) focused on errors associated with explaining away inferences in causal reasoning. They showed that people tend to be more accurate when they experience situations for which they are drawing causal inferences compared to situations that are simply described. Results suggest that adherence to normative causal reasoning depends on how causal models are presented, whether they are described or experienced directly.

Another bias that people exhibit when reasoning about competing causes is the zero-sum fallacy. Zero-sum reasoning broadly represents thinking where gains in one area take place at the expense of another's losses. In the context of causal reasoning, this is represented by treating evidence in support of a given cause as evidence against an alternative cause. In a recent study by Pilditch et al. (2019), people displayed a zero-sum bias when interpreting competing causes. When evidence was equally predicted by two competing causes, it was treated as irrelevant and as a result, was disregarded.

A balanced evaluation of evidence in legal cases involves weighing up evidence against competing hypotheses. These hypotheses are often about the causes that lead to outcomes under examination. Making accurate inferences requires not only correct interpretation of the weight of evidence, but also being able to correctly identify the hypotheses against which evidence is tested. Hypotheses can be considered mutually exclusive and exhaustive only when one (and only one) of the hypotheses can be true, ruling out any other explanation. For example, someone either dies from natural or unnatural causes. However, evidence in reality rarely warrants exclusivity and exhaustiveness of causes. There are usually many unknown possible causes of any piece of evidence. Being able to differentiate hypotheses that are not mutually exclusive and exhaustive is critical to avoiding the zero-sum fallacy, which occurs when hypotheses that are not mutually exclusive and exhaustive are erroneously treated as such.

Diagnostic and Predictive Causal Reasoning

Inferences from causes to effects represent predictive reasoning and moving from effects to causes corresponds to diagnostic reasoning. In a study of diagnostic causal reasoning with verbal probabilistic expressions, such as "frequently," "rarely," "likely" and "probably, " Meder and Mayrhofer (2017) found that inferences based on qualitative verbal terms, which are more widely used in everyday life to express uncertainty than numerical expressions, match those that are drawn from numerical information only. Overall, the study provided support for the human ability to make accurate probabilistic judgments, closely aligned with normative standards of Bayesian causal reasoning.

Diagnostic reasoning is underpinned not only by probabilistic judgments about cause given effect, but also by causal relations that connect causes to effects (Meder et al., 2014). The plausibility of causal models, in particular, is seen as one of the key factors impacting diagnostic judgments. According to this study, errors in observed diagnostic inferences can often be explained by variations in underlying causal models.

Hayes et al. (2018) suggested that the role of causal models in normative judgments merits further study. The authors were interested in assessing whether representations of causal models facilitate Bayesian probabilistic judgments in terms of normative accuracy as well as reduction in error magnitude. Participants were provided with causal explanations for statistical information (e.g., false positives) and their judgments for the likelihood of the corresponding events were compared with normative standards. The study results suggest that while providing causal explanations does not result in improved normative judgments, it can still help alter people's causal models by drawing attention to the statistical information which gets incorporated into causal structures.

Are Jurors Intuitive Bayesian Statisticians?

While the normative interpretation of the Bayesian formula implies that beliefs about guilt will be updated based on evidence, the judgments may not always match the quantitative Bayesian norms even when the qualitative interpretation is accurate.

Previous research suggests that jurors are competent at evaluating scientific evidence but tend to show systematic errors in processing quantitative evidence under certain circumstances (Hans et al., 2011). The discrepancy between the observed and quantitative normative updating judgments increases with the amount of evidence (Schum, 1966). One cause for this discrepancy may be the increased difficulty of estimating the diagnosticity of the available evidence when it is expressed in high numerical values. Dartnall and Goodman-Delahunty (2006) claimed that people are not sensitive to the probative weight of the probabilistic evidence. Such empirical evidence prompted researchers to see jurors as incompetent in intuitive probabilistic reasoning, prone to errors and systematic violation of rational belief updating principles (Arkes and Mellers, 2002).

Thompson et al. (2013) criticize the claims that people are always conservative Bayesian thinkers; instead they provide evidence that belief updating in relation to the quantitative evidence in criminal cases is in line with Bayesian norms. The authors argue that methodological limitations in earlier studies may have resulted in inferring that deviations from Bayesian norms in participants' observed judgments were more conservative than they actually were. Drawing on measures that were designed to address methodological shortcomings of previous studies, Thompson et al. (2013) found that while people at times engage in erroneous statistical reasoning, this is not always the case and people often reason in line with Bayesian belief updating models.

Present Study

The present study explores how people update beliefs in light of evidence, examining alignment with Bayesian norms from a qualitative (direction of updating) as well as a quantitative (numeric computations) perspective. The study focuses on the following aspects of causal reasoning: (1) predictive inferences from effects to causes; (2) diagnostic inferences from causes to effects; and (3) explaining away inferences with competing independent causes.

The present study is based on a summary of a real case where a couple was accused of intentionally harming their baby. In this case, a young child was brought to hospital by his parents because they noticed the child had bleeding in his mouth. The parents had no explanation for the bleeding, and said that the child had not been involved in an accident. In our experiment participants are given two possible causes for the bleeding: abuse and a rare blood disorder.

Participants are provided with information about the hospital admission rates for children with this symptom for cases of abuse and rare blood disorder. The story mentioned that figures from previous hospital admissions suggest that 1 in 100 children admitted with bleeding to the mouth have been abused by their parents, and 1 in 1,000 have the rare blood disorder.

After presenting background information, further evidence was presented one piece at a time. This involved information about:

(1) Doctors noticing bruising on the child
(2) The hospital radiologist carrying out an X-ray on the child and reporting that the X-ray showed fractures.
(3) The child being tested for the blood disorder and testing positive.
(4) An independent expert radiologist employed by the prosecution re-examining the X-ray results and claiming there were no fractures.

The causal structure of the case is presented in **Figure 1**.

EXPERIMENT 1

The main goal of Experiment 1 was to examine whether participants' beliefs are updated in line with Bayesian norms when dealing with competing causes (abuse and blood disorder) in a sequential inference task. Evidence was presented in stages, one piece of evidence at a time.

Method

Participants

155 participants were recruited through Amazon Mechanical Turk to take part in the study. In all experiments, participation was restricted to respondents who had at least a 95% approval rating for their previous MTurk work. Participants were English speakers and based in the United States. Participants who were unable to correctly answer the comprehension check questions regarding the underlying causal structure, were excluded from the analysis, leaving 127 participants (49 female). The mean age was 33.9 (SD = 11.04, range 73–13 = 55). Out of the 127 participants included in the study, 36.2% had an undergraduate degree, 10% – Masters or Ph.D. degree, 32.3% completed a college education and 22% had no qualification.

Design and Procedure

In the present study base rates are presented in a frequency format as research (Hoffrage et al., 2015; Woike et al., 2017; Weber et al., 2018) suggests that natural frequencies are preferred over probabilities in Bayesian reasoning tasks to minimize errors in inferences. When presented with background information about the case, participants were told that "Figures from previous

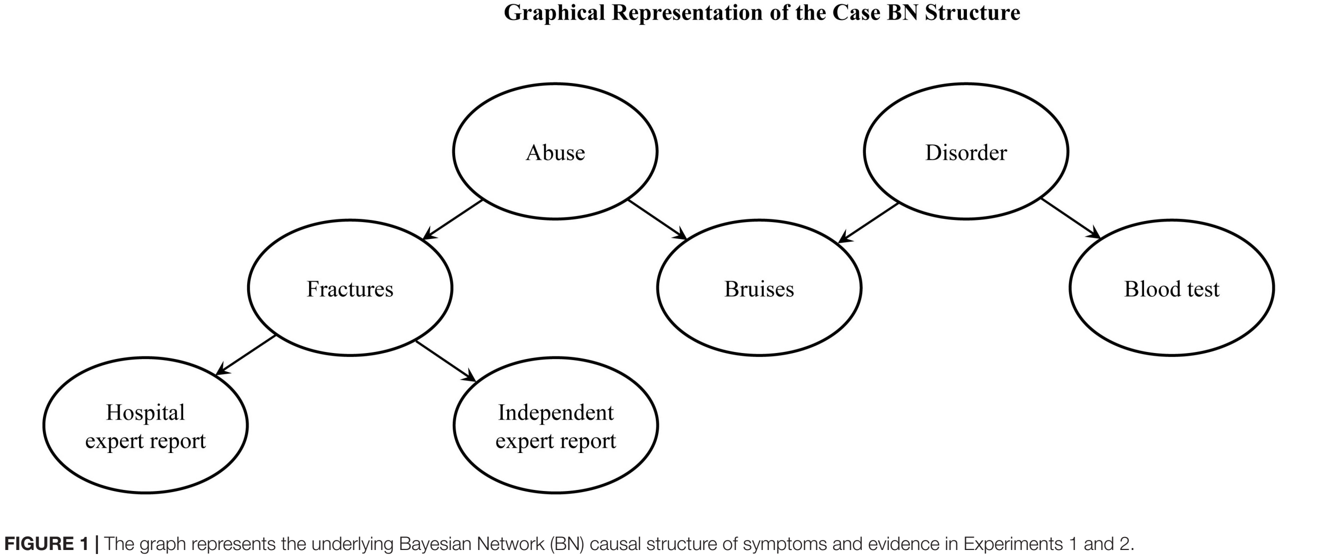

Graphical Representation of the Case BN Structure
Abuse
Disorder
Fractures
Bruises
Blood test
Hospital expert report
Independent expert report

FIGURE 1 | The graph represents the underlying Bayesian Network (BN) causal structure of symptoms and evidence in Experiments 1 and 2.

TABLE 1 | Task formulation.

In Experiments 1 and 2	Text	Responses
Introduction	A young child was brought to hospital by his parents because they noticed the child had bleeding in his mouth. The parents had no explanation for the bleeding, and said the child had not been involved in an accident. Doctors suggested two possible causes for the bleeding: abuse and a rare blood disorder.	
Statistical information	Figures from previous hospital admissions suggest that 1 in 100 children admitted with bleeding to the mouth have been abused by their parents, and 1 in 1000 have the rare blood disorder.	When responding to questions about base rates, this information remained visible to participants.
Questions after introduction and each stage of evidence presentation	• What are the chances that the parents abused the child? • What are the chances that the child has the blood disorder?	• Responses on a scale of 0% to 100%
Conditional probability questions showing probability of an event given the occurrence of other event(s)	**Questions about the bruises** • If the child has been abused but does NOT have the blood disorder, how likely is he to have bruises? • If the child has NOT been abused but does have the blood disorder, how likely is he to have bruises? • If the child has been abused and also has the blood disorder, how likely is he to have bruises? • If the child has NOT been abused and does NOT have the blood disorder, how likely is he to have bruises? **Questions about the blood test** • If the child has the blood disorder, how likely is he to test positive? • If the child does NOT have the blood disorder, how likely is he to test positive? **Questions about the fractures** • If the child has been abused, how likely is he to have fractures? • If the child has NOT been abused, how likely is he to have fractures? **Questions about the hospital radiologist report** • If the child has fractures, how likely is the hospital radiologist to report that he has fractures? • If the child does NOT have fractures, how likely is the hospital radiologist to report that he has fractures? **Questions about the independent radiologist report** • If the child has fractures, how likely is the expert radiologist to report that he has fractures? • If the child does NOT have fractures, how likely is the expert radiologist to report that he has fractures?	• Responses on a scale of 0 to 100 where "o" = Very unlikely, "100" = Very likely)

hospital admissions suggest that 1 in 100 children admitted with bleeding to the mouth have been abused by their parents, and 1 in 1,000 have the rare blood disorder."

The following evidence was presented in stages: bruises, a hospital x-ray expert's report, blood test results and an independent x-ray expert's report. To examine the possibility of zero-sum reasoning when assessing evidence (Pilditch et al., 2019) we varied the instructions given to participants about the exclusivity and exhaustiveness of the causes (abuse and blood disorder).

Participants were divided into three groups according to the presentation format for the abuse and causes. The experiment consisted of the following conditions:

Condition 1: Abuse and disorder were presented as non-exclusive causes of the child's bruises and bleeding.

Condition 2: Abuse and disorder were presented as non-exclusive and non-exhaustive causes of the child's bruises and bleeding.

Condition 3: Control condition contained no statement about the relationship between abuse and disorder as causes of the child's bruises and bleeding.

The dependent measures included the probabilistic judgments about the abuse and disorder as causes of the child's symptoms. The probability judgments were recorded after introducing the background information as well as after exposure to each new element of evidence (bruising, a hospital radiologist's report, blood test results and an independent radiologist's report).

Information presented to participants specified that bruising was a common consequence of abuse and also of the blood disorder. It further stated that fractures were a common consequence of abuse, but not of the blood disorder.

Conditional probabilities elicited from the participants at the end of the experiment were used to construct their models of evidence evaluation based on Bayesian reasoning. Individual Bayesian network models were constructed for each participant. Actual probability judgments were compared to those predicted by these models. The differences between the probabilistic judgments predicted by the subjective models in line with Bayesian norms and the actual probabilistic judgments formed a dependent measure in this experiment. Probability judgments were compared to the individual causal models inferred from conditional probabilities. Subjective priors were compared to base rates supplied in the introduction.

The probability judgments in the experiment were recorded on a scale of 0 to 100%. Participants used an on-screen slider with numerical values to indicate their answer. Questions about the probability judgments used the following format: "What are the chances of . . .?". On the slider response scale, 0% was labeled as "Very unlikely" and 100% as "Very likely".

The magnitude of updating from one stage of evidence to another was calculated as a difference between the probability estimates at the present and previous evidence stage, at Stage 2 (Bruises), Stage 3 (Hospital expert report), Stage 4 (Blood test results), and Stage 5 (Independent expert report).

The experiment was hosted on Qualtrics[1]. Participants were given a legal case (see **Table 1**). After reading the background information which contained priors for the probability of abuse and disorder as possible causes for the child's symptoms, participants were presented with four pieces of evidence in stages, one piece at a time. Starting from the introduction of the case background, participants were asked to provide their probability estimates for the abuse and the disorder as possible causes separately ("What are the chances that the parents abused the child?," "What are the chances that the child has the blood disorder?"). This process was followed throughout the experiment, eliciting subjective probabilities for abuse and disorder every time new evidence was presented. The order of questions was fixed and followed the sequence of evidence presentation. Conditional probabilities were also elicited after all pieces of evidence were presented and included questions such as "If the child has the blood disorder, how likely is he to test positive?," "If the child has fractures, how likely is the hospital radiologist to report that he has fractures?"). The procedure was adopted to track belief revision alongside the introduction of new evidence.

Results

The effect of Evidence and Condition on the abuse probability judgments in the observed data was examined with a mixed ANOVA with Condition as a between-subject and Evidence Stage as a within-subject variable. Following a Greenhouse-Geisser correction, the main effect of evidence was statistically significant, $F(2.545, 315.587) = 85.298$, $p < 0.001$, and partial eta squared = 0.408. Pairwise comparisons indicated that there was a statistically significant shift in beliefs about abuse at each stage of evidence, suggesting that participants integrated evidence and revised beliefs following the presentation of evidence. There was no main effect of Condition, $F(2, 124) = 1.201$, $p = 0.304$.

A mixed ANOVA was carried out to explore the effect of Evidence Stage and Condition on the observed subjective probability judgments for disorder. Similar to the abuse probability judgments, a significant main effect of Evidence was found following a Greenhouse-Geisser correction, F(2.293, 284.339) = 479.268, $p < 0.001$, partial eta squared = 0.794.

[1] www.qualtrics.com

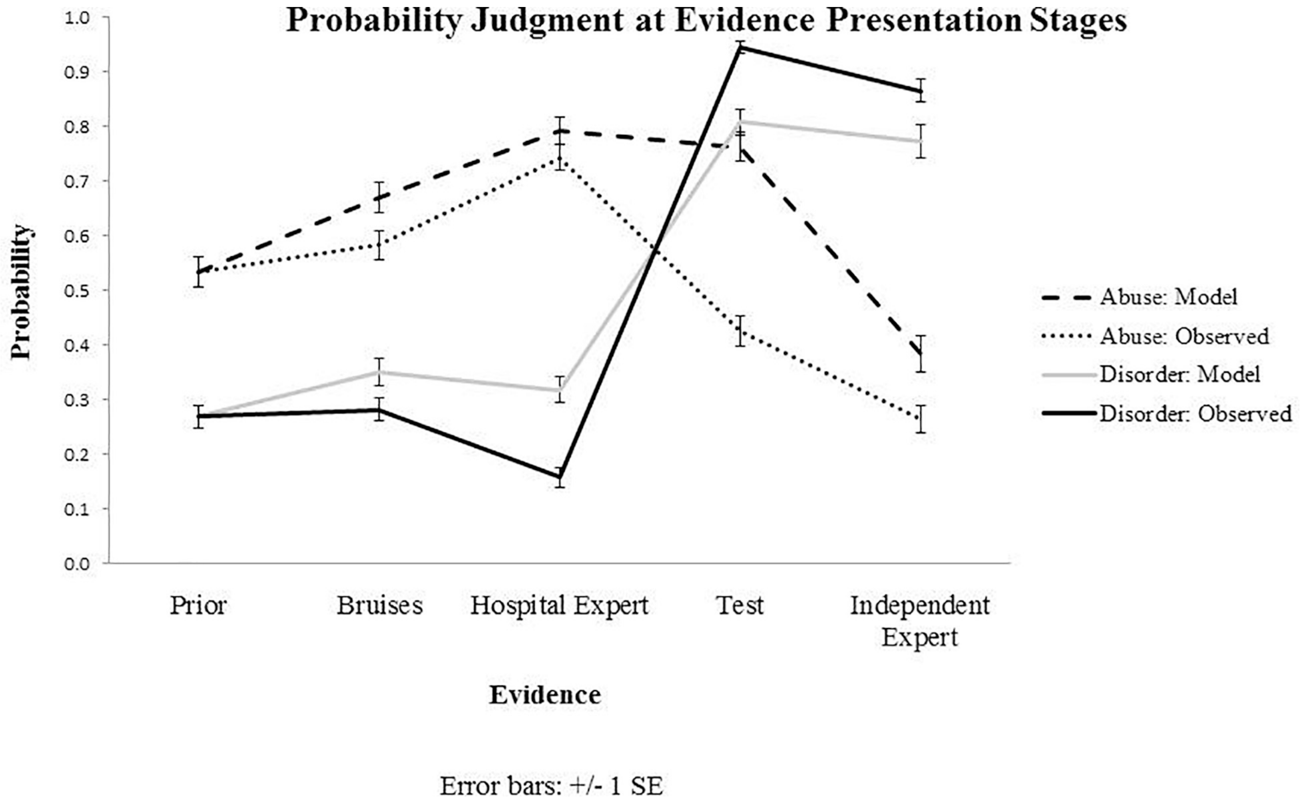


FIGURE 2 | Results from Experiment 1: Observed and predicted (model) probability judgments at each evidence presentation stage, starting with prior beliefs and capturing belief updating following the evidence about bruises, hospital expert report, blood test results and independent expert report. Priors presented on the graph for both observed and predicted values represent subjective priors set by the participants.

There was no effect of Condition, $F(2, 124) = 0.127$, $p = 0.881$. *Post hoc* comparisons using the Bonferroni test indicated no difference between prior beliefs ($M = 0.269$, SD = 0.021) and revised beliefs after evidence about bruises ($M = 0.282$, SD = 0.022). All other stages of belief updating, including the first expert's report ($M = 0.157$, SD = 0.018), blood test results ($M = 0.945$, SD = 0.011) and the second expert's report ($M = 0.867$, SD = 0.022) showed differences in beliefs compared to the previous stage. Subjective priors were considerably higher ($M_{abuse} = 0.534$, SD = 0.305; $M_{disorder} = 0.267$, SD = 0.238) than the objective priors (0.01 and 0.001, respectively) supplied as part of the case scenario.

Individual Bayesian belief updating models were obtained using the gRain package in R (Højsgaard, 2012). Differences between the observed and predicted (Bayesian) probability judgments during the belief updating process are summarized in **Figures 2**, **3**, which draw on the participants' own priors.

An example Bayesian belief updating model is presented in **Figure 4**.

Discussion

Differences between the observed and predicted judgments were found to be significant for all pieces of evidence, including bruising symptoms, hospital radiologist's report and test results with regards to the abuse and disorder-related probability judgments indicating that the participants' probabilistic judgments were different from exact Bayes computations. However, judgments were qualitatively in the right direction. There was no significant difference between conditions in belief updating, indicating that making the non-exclusivity and non-exhaustiveness of causes explicit did not affect probability judgments.

EXPERIMENT 2

Experiment 2 focused on testing belief updating when participants were explicitly told, just before each probability judgment, that the causes in the study were non-exclusive and non-exhaustive. The purpose of this experiment was to determine the effect of bringing participants' attention to

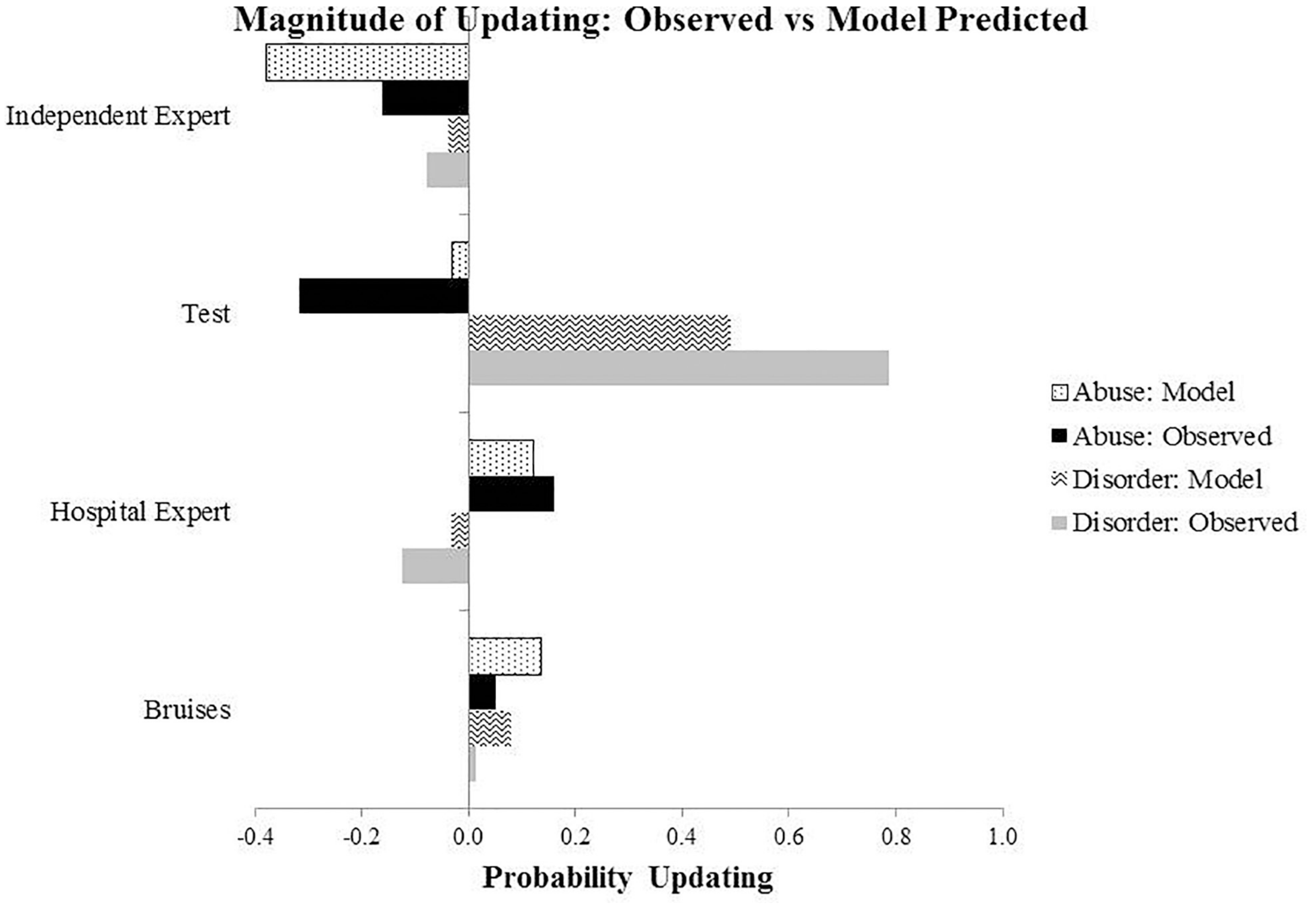
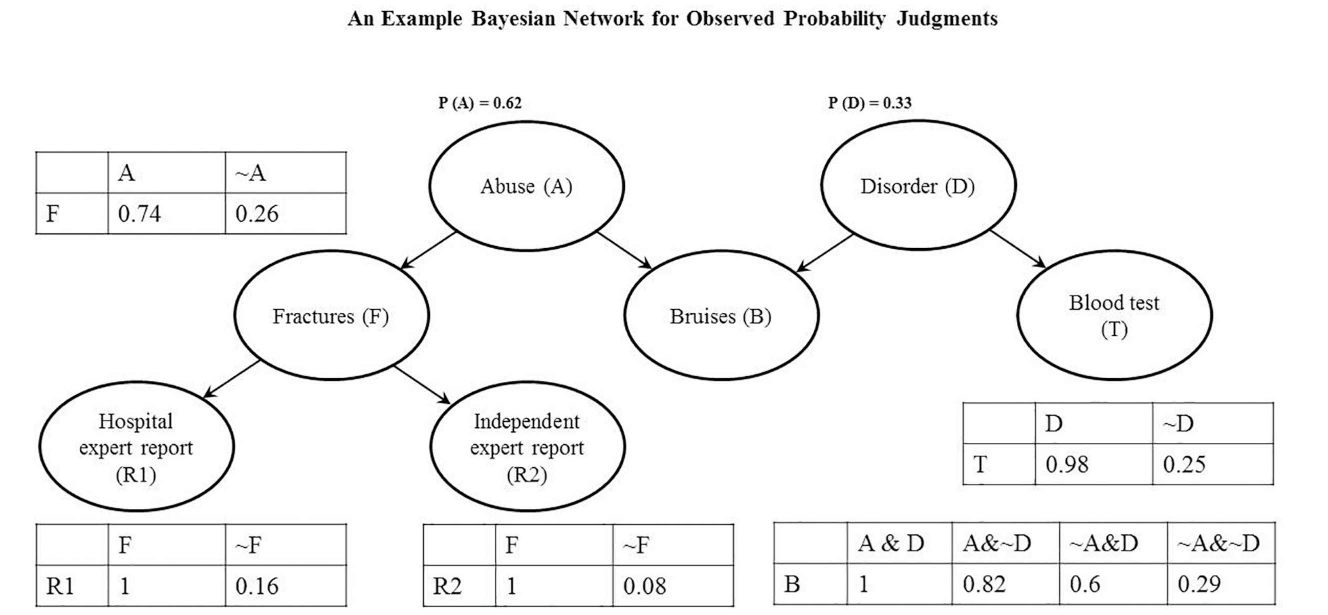


FIGURE 3 | Results from Experiment 1: The graph shows extent of updating at four stages of evidence presentation, calculated as a difference between the probability estimates at the present and previous evidence stage. For example, for the Bruises stage, this is calculated by subtracting the probability value at the previous stage (prior elicitation) from the present stage (evidence of bruises).


An Example Bayesian Network for Observed Probability Judgments
P (A) = 0.62
P (D) = 0.33
Abuse (A)
Disorder (D)
Fractures (F)
Bruises (B)
Blood test (T)
Hospital expert report (R1)
Independent expert report (R2)

	A	~A
F	0.74	0.26

	D	~D
T	0.98	0.25

	F	~F
R1	1	0.16

	F	~F
R2	1	0.08

	A & D	A&~D	~A&D	~A&~D
B	1	0.82	0.6	0.29

FIGURE 4 | Individual Bayesian Network: The graph shows a Causal Bayesian Network and corresponding probability tables for one of the participants from Experiment 1.

the non-exclusivity and non-exhaustiveness of causes on the accuracy of judgments. This would allow us to rule out lack of understanding of the causal structure as a contributing factor to biased judgments observed in Experiment 1. The following statement was included at every stage of subjective probability elicitation: "Note that it is possible that both causes are true: e.g., that a child has been abused and has the disorder; it is also possible that neither are true, and that the symptoms arise due to other causes." Additionally, participants' understanding of the case causal structure was tested.

Method

Participants

93 participants were recruited using the same protocol as in Experiment 1. As in Experiment 1, participation was restricted to respondents who had at least a 95% approval rating for their previous MTurk work. Participants were English speakers and based in the United States. The mean age was 35.08 (SD = 12.64, range 74–19 = 55). Out of the 93 participants (52 female) included in the study, 44% had an undergraduate degree, 21.5% – Masters or PhD degree and 23.7% completed a college education.

Design and Procedure

The procedure, instruction, and materials, including the questions were identical to those used in Experiment 1 except there was only one Condition, which corresponded to Condition 2 in Experiment 1. Additionally, at the end of the task we included questions to elicit participants' causal models, focusing on the links included in the case model (**Figure 1**). Questions followed the format: "Did A cause B?"

Results

A repeated measures ANOVA with a Greenhouse-Geisser correction showed that mean probability estimates differed significantly between the evidence presentation stages [$F(2.631, 194.728) = 300.192$, $p < 0.001$, partial eta squared = 0.802], observed and model judgments based on Bayesian predictions [$F(1, 74) = 45.09$, $p < 0.001$, partial eta squared = 0.379], but not between the abuse and disorder probability judgments [$F(1, 74) = 1.334$, $p = 0.252$].

Subjective priors were higher ($M_{abuse} = 0.467$, SD = 0.313; $M_{disorder} = 0.441$, SD = 0.325) than the objective priors (0.01 and 0.001, respectively) supplied as part of the case scenario.

Belief updating, drawing on the participants' own subjective priors, is summarized in **Figures 5**, **6**.

To check for the general accuracy of the underlying causal models, we tested for the links that were included in the causal structure of the case model (**Figure 1**) as well as links that were incompatible with the model. Results for the links that were used in Experiment 1 to test comprehension of causal models as a basis for screening participants who did not answer the questions correctly, showed that the accuracy of responses was above the chance level. Participants correctly identified the causal structure. Specifically, participants were able to correctly identify that bruising was a common consequence of abuse and also of the blood disorder and that fractures were a common consequence of abuse, but not of the blood disorder. In response to questions whether a certain causal link was present or not, participants were able to correctly identify that Abuse could cause Bruises above the chance level, i.e., above 0.5 (0.53), Disorder could cause Bruises (0.78), and Abuse could cause Fractures (0.52). With regard to a causal link that was not part of the case causal structure such as

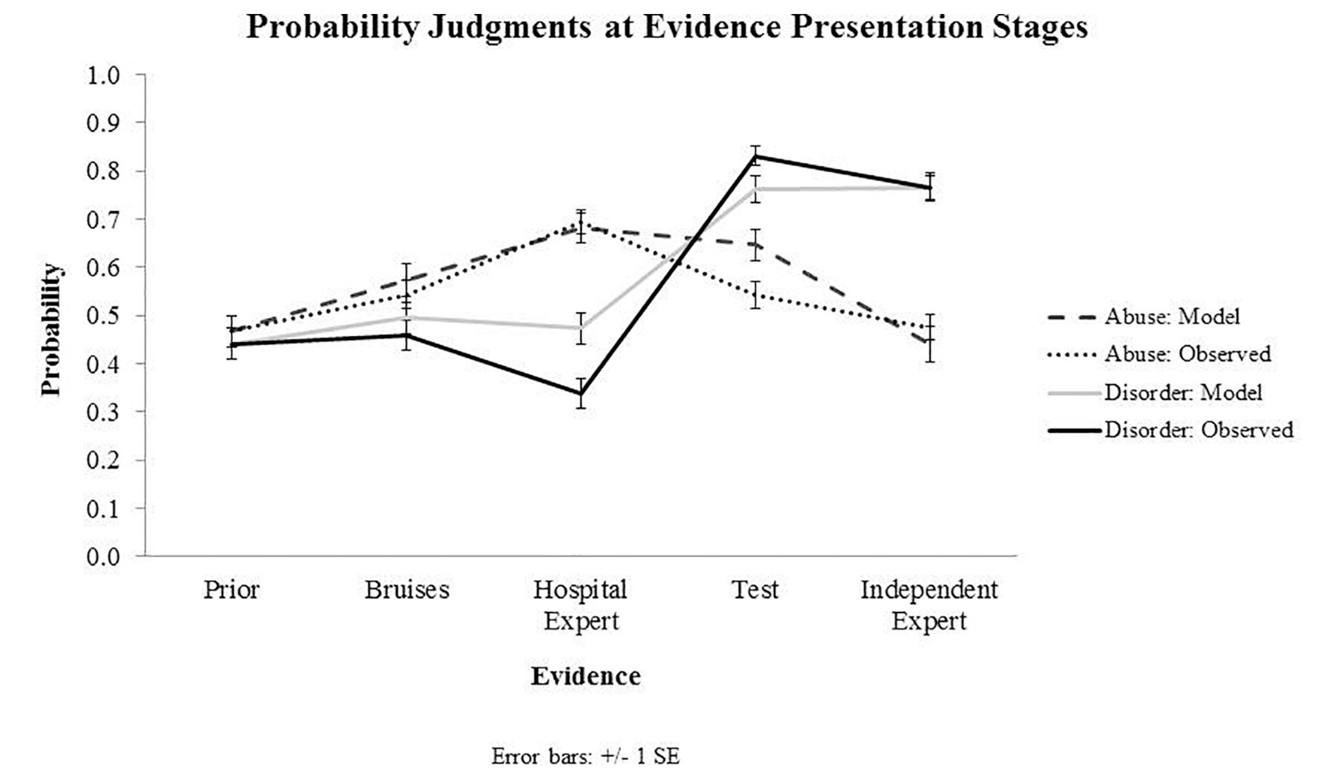


FIGURE 5 | Results from Experiment 2: The graph presents observed and predicted (model) probability judgments for Abuse and Disorder at each sequential stage of evidence presentation. Priors presented on the graph for both observed and predicted values represent subjective priors set by the participants.

Disorder causing Fractures, participants' responses showed that participants were able to correctly exclude this link from their individual causal representations (0.27).

Discussion

Participants updated beliefs in the direction predicted by normative Bayesian judgments on most occasions. However, as in Experiment 1, we found instances of under- and over-estimation of evidence in quantitative belief updating. This was particularly evident when integrating evidence that supported both Abuse and Disorder as possible causes (e.g., evidence of bruises), which led to the under-weighting of evidence and belief updating far lower than mandated by Bayesian normative judgments. Another instance of inaccurate quantitative judgment was following the positive blood test results, which showed that participants attributed excessive weight to evidence.

GENERAL DISCUSSION

The findings from both experiments suggest that in legal decision making people qualitatively update their beliefs in line with Bayesian norms. In our experiments, participants' belief updating was qualitatively aligned with normative judgments, i.e., probability judgments increased or decreased in the same direction as predicted based on Bayesian norms. This was observed with both predictive inferences (e.g., increases in the probability of abuse increased the probability of fractures), diagnostic inferences (e.g., evidence of the hospital radiologist report raised the subjective probability of fractures) and explaining away inferences (e.g., evidence of positive test results raised the probability of the blood disorder and decreased the probability of abuse by explaining away). While most judgments fit with qualitative predictions of Bayesian models, an exception is observed at the final stage of evidence presentation where the subjective probability of blood disorder was lowered slightly rather than raised. This can be explained by exposure to conflicting expert reports, which may have decreased the perceived reliability of reports, resulting in a greater skepticism toward the blood test results.

Overall the results indicate that people's qualitative reasoning is mostly accurate and follows qualitative predictions of Bayesian models in predictive, diagnostic and explaining away inferences. These findings reinforce results from previous studies where Bayesian probabilistic reasoning was observed (e.g., Thompson and Newman, 2015).

Results from both experiments indicate that people tend to ignore the priors provided as part of the background case information and set their own subjective priors. The subjective priors in both experiments were significantly higher than the objective priors offered in the case summary. Prior knowledge and expectations, underlying causal models may have contributed to setting higher priors than suggested by base rates.

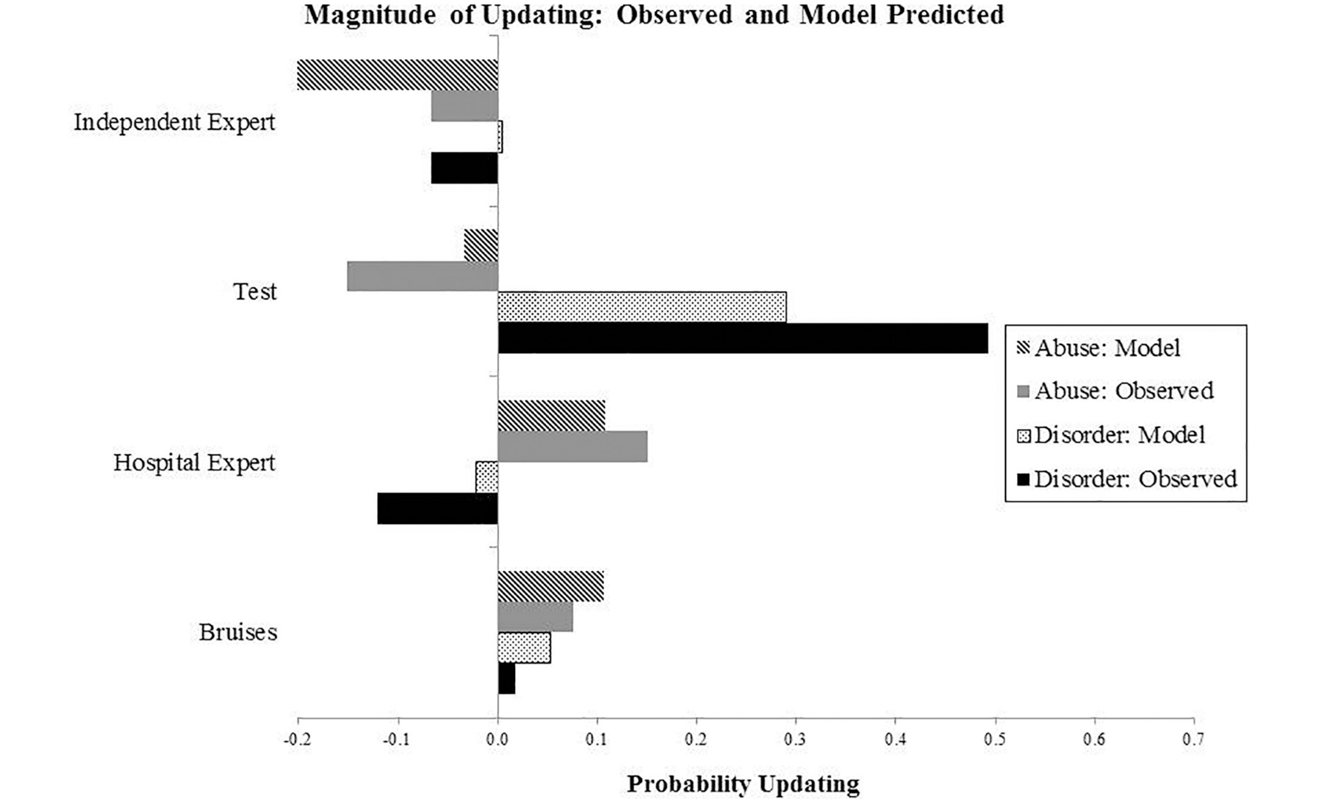


FIGURE 6 | Results from Experiment 2: The graph shows extent of updating at four stages of evidence presentation, calculated as a difference between the probability estimates at the present and previous evidence stage. For example, for the Hospital Expert evidence stage, this is calculated by subtracting the probability value at the previous stage (evidence of bruises) from the present stage (evidence of the hospital expert report).

Previous research about explaining away inferences is not conclusive (Morris and Larrick, 1995; Oppenheimer and Monin, 2009; Rehder, 2014; Rottman and Hastie, 2014, 2016; Tesic et al., 2020) and offers different views on challenges associated with explaining away inferences. Our experiments highlighted that people are able to navigate the explaining away type of scenarios and make accurate judgments about competing causes that fit with qualitative Bayesian predictions.

Prior evidence on people's ability to make quantitative probabilistic judgments aligned with Bayesian norms is not definitive, with studies indicating either under- or over-estimation of evidence. In the existing body of literature, a unified mechanism for explaining an excessively low and high weight attributed to evidence has not been decisively established. Our findings show both types of departures from quantitative normative judgments: under-weighting of evidence when participants update beliefs based on the evidence of bruises and over-weighting of evidence following the evidence of the positive test result. Both these findings can be explained with zero-sum reasoning, which provides insights into how people integrate evidence when dealing with competing causes (Pilditch et al., 2019).

Zero-sum reasoning represents thinking whereby the gains of one person take place at the expense of another's losses. A zero-sum model of the world presumes a finite and fixed amount of resources in the world, which necessitates a competition for these resources. In the context of competing causes, when two causes equally predict the same evidence, zero-sum thinking treats such evidence as neutral because it tacitly assumes that the causes are exclusive and exhaustive accounts of the evidence. For example, in our experiments the evidence of bruises which was predicted by both the abuse and disorder hypotheses, the evidence was treated as neutral, resulting in only slight increase in the probability of both, which was considerably lower than expected by Bayesian updating.

Zero-sum thinking also accounts for the over-weighting of evidence which was observed in the case of the excessive decrease in the probability of disorder following the hospital radiologist report and an excessive increase in the probability of disorder given the positive blood test result with a simultaneous disproportionate lowering of the probability of abuse. Excessive raising or lowering of probabilities points to the zero-sum nature of the reasoning involved in this process. Competing causes were perceived as exclusive, which had a hydraulic effect on the evidence interpretation: increasing the probability of one cause excessively decreased the probability of the other cause.

These findings are consistent with the results of Pilditch et al. (2019) who found that when interpreting evidence against competing causes, people treat evidence evaluation as a zero-sum game. The biased reasoning persisted even when the non-exhaustiveness of the hypotheses was made explicit. Our results also show that zero-sum thinking is observed despite the participants being made aware the non-exclusive and non-exhaustive nature of the competing causes.

Zero-sum reasoning in the context of our experiments suggests that under- and over-estimation of evidence are observed due to underlying assumptions about causes modeled on zero-sum principles. This type of reasoning may result in more accurate judgments when dealing with competing causes that are exclusive and exhaustive.

CONCLUSION

Our study suggests that people are able to make qualitatively accurate causal inferences and update beliefs in the direction predicted by Bayesian norms. However, quantitative computations are not always accurate and show a gap between observed and normative judgments. Instances of underweighting and overweighting of evidence in our experiments can be explained by a zero-sum fallacy. This offers a useful perspective for shedding light on evidence integration in legal cases, where a balanced evaluation of evidence often involves weighing up of the evidence against competing hypotheses. These hypotheses are often about the causes that lead to outcomes under examination. Making accurate inferences requires not only correct interpretation of the weight of evidence, but also being able to identify the hypotheses against which evidence is tested. Being able to differentiate hypotheses that are not mutually exclusive and exhaustive is critical to avoiding a zero-sum fallacy.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Department of Experimental Psychology, University College London. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by ESRC grant ES/H041796/1 Causality and Responsibility Judgments to DL.

REFERENCES

Arkes, H. R., and Mellers, B. A. (2002). Do Juries Meet Our Expectations? *Law Hum. Behav.* 26, 625–639. doi: 10.1023/a:1020929517312

Baetu, I., and Baker, A. G. (2009). Human Judgments of Positive and Negative Causal Chains. *J. Exp. Psychol. Anim. Behav. Processes* 35, 153–168. doi: 10.1037/a0013764

Chater, N., and Oaksford, M. (eds) (2008). *The probabilistic mind prospects for Bayesian cognitive science.* Oxford: Oxford University Press.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends Cognit. Sci.* 10, 287–291. doi: 10.1016/j.tics.2006.05.007

Dartnall, S., and Goodman-Delahunty, J. (2006). Enhancing Juror Understanding of Probabilistic DNA Evidence. *Aus. J. Forensic Sci.* 38, 85–96. doi: 10.1080/00450610609410635

Fenton, N., Neil, M., and Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognit. Sci.* 37, 61–102. doi: 10.1111/cogs.12004

Fenton, N., Neil, M., and Thieme, N. (2018). Lawnmowers versus terrorists. *Significance* 15, 12–13. doi: 10.1111/j.1740-9713.2018.01104.x

Hans, V., Kaye, D., Dann, B., Farley, E., and Albertson, S. (2011). Science in the Jury Box: Jurors' Comprehension of Mitochondrial DNA Evidence. *Law Hum. Behav.* 35, 60–71. doi: 10.1007/s10979-010-9222-8

Hayes, B., Ngo, J., Hawkins, G., and Newell, B. (2018). Causal explanation improves judgment under uncertainty, but rarely in a Bayesian way. *Mem. Cogn.* 46, 112–131. doi: 10.3758/s13421-017-0750-z

Hoffrage, U., Krauss, S., Martignon, L., and Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Front. Psychol.* 6:1473. doi: 10.3389/fpsyg.2015.01473

Højsgaard, S. (2012). Graphical Independence Networks with the gRain Package for R. *J. Statist. Softw.* 46, 1–26.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behav. Brain Sci.* 19, 1–17. doi: 10.1017/S0140525X00041157

Krynski, T. R., and Tenenbaum, J. B. (2007). The Role of Causality in Judgment Under Uncertainty. *J. Exp. Psychol. General* 136, 430–450. doi: 10.1037/0096-3445.136.3.430

Lagnado, D. A., and Gerstenberg, T. (2017). "Causation in Legal and Moral Reasoning," in *The Oxford Handbook of Causal Reasoning*, 1 Edn, ed. M. R. Waldmann (Oxford: Oxford University Press).

Liefgreen, A., Tesic, M., and Lagnado, D. (2018). "Explaining away: signi?cance of priors, diagnostic reasoning, and structural complexity," in *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, (Texas: Cognitive Science Society).

Meder, B., and Mayrhofer, R. (2017). Diagnostic causal reasoning with verbal information. *Cognit. Psychol.* 96, 54–84. doi: 10.1016/j.cogpsych.2017.05.002

Meder, B., Hagmayer, Y., and Waldmann, M. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bull. Rev.* 15, 75–80. doi: 10.3758/PBR.15.1.75

Meder, B., Mayrhofer, R., and Waldmann, M. R. (2014). Structure Induction in Diagnostic Causal Reasoning. *Psychol. Rev.* 121, 277–301. doi: 10.1037/a0035944

Morris, M. W., and Larrick, R. P. (1995). When One Cause Casts Doubt on Another: A Normative Analysis of Discounting in Causal Attribution. *Psychol. Rev.* 102, 331–355. doi: 10.1037/0033-295X.102.2.331

Oppenheimer, D., and Monin, B. (2009). Investigations in spontaneous discounting. *Memory Cognit.* 37, 608–614. doi: 10.3758/MC.37.5.608

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference/Judea Pearl*. San Mateo: Morgan Kaufmann.

Pearl, J. (2000). *Casuality : models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pennington, N., and Hastie, R. (1986). Evidence Evaluation in Complex Decision Making. *J. Personal. Soc. Psychol.* 51, 242–258. doi: 10.1037/0022-3514.51.2.242

Pennington, N., and Hastie, R. (1992). Explaining the Evidence: Tests of the Story Model for Juror Decision Making. *J. Personal. Soc. Psychol.* 62, 189–206. doi: 10.1037/0022-3514.62.2.189

Peterson, C. R., and Beach, L. R. (1967). Man as an Intuitive Statistician. *Psychol. Bull.* 68, 29–46. doi: 10.1037/h0024722

Phillips, L. D., and Edwards, W. (1966). Conservatism in a simple probability inference task. *J. Exp. Psychol.* 72, 346–354. doi: 10.1037/h0023653

Pilditch, T. D., Fenton, N., and Lagnado, D. (2019). The Zero-Sum Fallacy in Evidence Evaluation. *Psychol. Sci.* 30, 250–260. doi: 10.1177/0956797618818484

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognit. Psychol.* 72, 54–107. doi: 10.1016/j.cogpsych.2014.02.002

Rehder, B., and Waldmann, M. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Mem. Cogn.* 45, 245–260. doi: 10.3758/s13421-016-0662-3

Rottman, B. M. (2017). "The Acquisition and Use of Causal Structure Knowledge," in *The Oxford Handbook of Causal Reasoning*, 1 Edn, ed. M. R. Waldmann (Oxford: Oxford University Press).

Rottman, B. M., and Hastie, R. (2014). Reasoning About Causal Relationships: Inferences on Causal Networks. *Psychol. Bull.* 140, 109–139. doi: 10.1037/a0031903

Rottman, B. M., and Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognit. Psychol.* 87, 88–134. doi: 10.1016/j.cogpsych.2016.05.002

Schum, D. A. (1966). Prior uncertainty and amount of diagnostic evidence as variables in a probabilistic inference task. *Organiz. Behav. Hum. Perfor.* 1, 31–54. doi: 10.1016/0030-5073(66)90004-3

Sloman, S. (2005). Avoiding foolish consistency. *Behav. Brain Sci.* 28, 33–34. doi: 10.1017/S0140525X05430010

Sloman, S. A., and Lagnado, D. (2015). Causality in Thought. *Annu. Rev. Psychol.* 66, 223–247. doi: 10.1146/annurev-psych-010814-015135

Sloman, S., and Lagnado, D. (2005). Do We "do"? *Cognit. Sci.* 29, 5–39.

Tesic, M., Liefgreen, A., and Lagnado, D. (2020). The propensity interpretation of probability and diagnostic split in explaining away. *Cognit. Psychol.* 121, 101293–101293. doi: 10.1016/j.cogpsych.2020.101293

Thompson, W. C., and Newman, E. J. (2015). Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents. *Law Hum. Behav.* 39, 332–349. doi: 10.1037/lhb0000134

Thompson, W. C., Kaasa, S. O., and Peterson, T. (2013). Do Jurors Give Appropriate Weight to Forensic Identification Evidence? *J. Empirical Legal Stud.* 10, 359–397. doi: 10.1111/jels.12013

Tversky, A., and Kahneman, D. (1982). "Evidential impact of base rates," in *Judgment under uncertainty: Heuristics and biases*, eds D. Kahneman, P. Slovic, and A. Tversky (New York: Cambridge University Press), 153–160. doi: 10.1017/cbo9780511809477.011

Waldmann, M. R. (2000). Competition Among Causes But Not Effects in Predictive and Diagnostic Learning. *J. Exp. Psychol. Learning Memory Cognit.* 26, 53–76. doi: 10.1037/0278-7393.26.1.53

Weber, P., Binder, K., and Krauss, S. (2018). Why Can Only 24% Solve Bayesian Reasoning Problems in Natural Frequencies: Frequency Phobia in Spite of Probability Blindness. *Front. Psychol.* 9:1833. doi: 10.3389/fpsyg.2018.01833

Wilson, W. (2017). *Criminal law*. Harlow: Pearson Education.

Woike, J. K., Hoffrage, U., and Martignon, L. (2017). Integrating and testing natural frequencies, naïve Bayes, and fast-and-frugal trees. *Decision* 4, 234–260. doi: 10.1037/dec0000086

Conflict of Interest: 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 handling editor declared a shared affiliation, though no other collaboration, with the authors at the time of the review.

ORIGINAL RESEARCH
published: 05 March 2021
doi: 10.3389/fpsyg.2021.525195

The Development of Spatial–Temporal, Probability, and Covariation Information to Infer Continuous Causal Processes

Selma Dündar-Coecke[1]*, *Andrew Tolmie*[1] *and Anne Schlottmann*[2]

[1] Centre for Educational Neuroscience and Department of Psychology and Human Development, UCL Institute of Education, University College London, London, United Kingdom, [2] Department of Experimental Psychology, University College London, London, United Kingdom

This paper considers how 5- to 11-year-olds' verbal reasoning about the causality underlying extended, dynamic natural processes links to various facets of their statistical thinking. Such continuous processes typically do not provide perceptually distinct causes and effect, and previous work suggests that spatial–temporal analysis, the ability to analyze spatial configurations that change over time, is a crucial predictor of reasoning about causal mechanism in such situations. Work in the Humean tradition to causality has long emphasized on the importance of statistical thinking for inferring causal links between distinct cause and effect events, but here we assess whether this is also viable for causal thinking about continuous processes. Controlling for verbal and non-verbal ability, two studies (N = 107; N = 124) administered a battery of covariation, probability, spatial–temporal, and causal measures. Results indicated that spatial–temporal analysis was the best predictor of causal thinking across both studies, but statistical thinking supported and informed spatial–temporal analysis: covariation assessment potentially assists with the identification of variables, while simple probability judgment potentially assists with thinking about unseen mechanisms. We conclude that the ability to find out patterns in data is even more widely important for causal analysis than commonly assumed, from childhood, having a role to play not just when causally linking already distinct events but also when analyzing the causal process underlying extended dynamic events without perceptually distinct components.

Keywords: probability, covariation, spatial–temporal thinking, causation, causal processes, development

OPEN ACCESS

Edited by:
Andrea Bender,
University of Bergen, Norway

Reviewed by:
Christoph Hoerl,
University of Warwick,
United Kingdom
Nicole Cruz,
University of New South Wales,
Australia

****Correspondence:***
Selma Dündar-Coecke
selma.coecke@gmail.com
orcid.org/0000-0003-2987-170X

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

Received: *08 January 2020*
Accepted: *27 January 2021*
Published: *05 March 2021*

Citation:
Dündar-Coecke S, Tolmie A and Schlottmann A (2021) The Development of Spatial–Temporal, Probability, and Covariation Information to Infer Continuous Causal Processes.
Front. Psychol. 12:525195.
doi: 10.3389/fpsyg.2021.525195

INTRODUCTION

Hume (1739/1978) argued that we can only know about causality from the "constant conjunction" of potential causes and effects. Since then, multiple schools of thought have put some form of statistical analysis of repeated experience at the core of causal thinking, ranging from the causal attribution literature in social psychology (Kelley, 1967, 1973) to work on associative causal learning inspired by animal studies (Shanks and Dickinson, 1988). However, although causes covary with their effects, inferring causation from correlation has many pitfalls. Kantian reasoning instead focuses on the underlying causal mechanisms that allow causes to generate their effects, and modern approaches attempt to integrate such mechanism-based thinking with statistical analyses (see Waldmann, 2017).

When given a choice, people tend to seek information about mechanisms (how a process works) rather than covariation (inferring joint variability of two random variables) to determine causality (Ahn et al., 1995). People seem to recognize that statistical information needs to fit with the mechanism, because it is the latter that generates the covariation of cause and effect. However, in many situations, the underlying generative mechanism is unknown. In such cases, statistical reasoning, forms of analyses based on information about the frequency of occurrence or co-occurrence of potential causes and effects, is crucial for causal thinking (e.g., Cheng, 1997; Griffiths and Tenenbaum, 2009).

Analyses of statistical regularities between events presuppose that there are separate events to be linked into cause-and-effect sequences, for instance, when pushing a button is followed by a light coming on or when a ball is set in motion by collision with another ball. Most studies of causal thinking have considered causal sequences with such distinct components.

We do, however, also reason about causality in naturally continuous processes, without clear segmentation into potential cause and effect, as when an object sinks, for instance, or dissolves in water. The observation here is of continuous change, and while we may think about what causes this change, or what causes one of its features, for instance, why one object sinks slow, while another sinks fast, in our perceptual experience, the process has no naturally distinct components to serve as candidate cause and effects. One can nevertheless focus, in thought, on aspects of the process and think about the underlying causal mechanism, of course, but it is not so clear anymore whether and how statistical reasoning contributes to causal reasoning here.

We have recently begun to study children's causal thinking in these types of dynamic natural processes, taking an individual difference approach and finding that measures of what we call spatial–temporal analysis were important predictors of children's thinking about the causal mechanisms involved (Dündar-Coecke et al., 2019, 2020). Spatial–temporal analysis is the ability to think about how spatial configurations change over time. It is separate from children's verbal and non-verbal IQ and from their spatial ability, which is not predictive of causal thinking. Spatial–temporal analysis goes beyond purely spatial analysis in that it includes the ability to work out the temporal order of a series of spatial states and the ability to project these state transformations onto past, present, and future experiences. Spatial–temporal analysis might thus help children find segmentations for continuous processes, which in turn would be helpful for causal reasoning about them.

In the present paper, we use a similar individual difference approach to return to the more Humean question of whether aspects of children's statistical thinking also predict their causal reasoning about continuous processes, and how such statistical predictors compare to their spatial–temporal predictors. We present further data from the project reported in Dündar-Coecke et al. (2019, 2020), which developed a set of novel tasks to look at children's causal thinking about continuous processes (sinking, absorption, and dissolving). It also involved a large battery of spatial–temporal, spatial, verbal, and non-verbal reasoning tasks, as previously reported, and in addition, the set of statistical reasoning tasks that are the focus of the present paper. In subsequent sections, we discuss in more detail our statistical thinking measures and their possible links to causal thinking.

ON THE LINK BETWEEN PROBABILISTIC THINKING AND CAUSAL PROCESSES

Probabilistic reasoning enables one to estimate of the likelihood of an event that may or may not occur (mud suggests rain). In a world where causal processes are induced by complex set of factors, it is crucial to analyze the degree of certainty of causal relationships because in most circumstances there are unobserved latent factors, which allow exceptions (not all mud suggests rain, but sometimes flooding). In some circumstances, probabilistic thinking can be used as a tool to improve the accuracy of our decisions even in the absence of mechanism knowledge.

Interest in the role of probability has already led to psychological investigation. In Piaget and Inhelder (1975) studies, the development of such thinking was seen as a formal operational achievement. The emphasis in this approach was on improvements in children's ability to quantify the relative proportions of target and non-target events as they get older. A more recent approach, in contrast, has focused on children's intuitive understanding. Multiple studies have shown that children's probability judgments conform to the structure of normative probability concepts, e.g., taking an appropriate ratio from kindergarten age (e.g., Anderson and Schlottmann, 1991; Schlottmann and Anderson, 1994; Acredolo et al., 1989; Schlottmann, 2001, reviewed in Schlottmann and Wilkening, 2012). Even younger, pre-school children already have a basic ability to discriminate predictable from unpredictable event sequence (Kuzmak and Gelman, 1986), and there have been multiple demonstrations in recent years that infants have some sensitivity to different sampling processes (Xu and Denison, 2009). Thus, early capabilities of engaging in rudimentary probability calculations co-exist with difficult tasks that are computationally challenging for young children and have high demand on memory skills (e.g., White, 2014; McCormack et al., 2015, 2016).

These demonstrations involve elaborate lengthy experimental tasks that would not be suitable for a correlational study. Here, we use the abbreviated versions of the probability tasks and investigate whether probabilistic thinking is relevant to reasoning of continuous causal processes. We hypothesize that children's ability to judge probability may not just index computational ability, but also sensitivity to definiteness of outcomes in the world. To test this hypothesis, we first observe children's sensitivity to probability along with their computational abilities. Further, we investigate whether the development of probability understanding is linked to children's reasoning about continuous causal phenomena (sinking, absorption, and solution). Third, we compare these competences with children's performances on spatial–temporal measures. The predictive tasks – probability, covariation, spatial–temporal – were presented to elicit whether children's computational ability or sensitivity to probability mattered for the inference of causal processes. This three-stage investigation helps us to identify how individual differences in such probability judgments might link with individual differences in reasoning about temporally extended processes above/beyond other reasoning types.

The tasks in which children exhibit these abilities typically involve non-causal models, displaying all outcome possibilities simultaneously to minimize memory requirements. For example, in the first probability task (marbles), the child sees a plate with seven red winner marbles and three blue loser marbles and judges how easy it is to win in a blind draw. In tasks where probabilities are experienced sequentially (e.g., the child draws a number of times from a population with initially unknown proportion of winner and loser marbles), children do not do so well when predicting the next outcome, as has long been known from work on probability learning (Brainerd, 1981) and child variants of the Iowa Gambling Task (Huizenga et al., 2007). Children's difficulties in sequential tasks may reflect memory capacity and other processing limitations, though, and in any case indicate problems with cumulative estimation rather than basic grasp of probability [these two types of tasks address different aspects of understanding, as discussed in Schlottmann and Wilkening's (2012) review].

The marbles task captures children's sensitivity to probability rather than their computational ability. It derives appreciation of uncertainty and likelihood from rational analysis that multiple outcomes are possible in a given situation and from enumeration of these outcomes, prior to experiencing instances of the outcomes themselves. Probability tasks laying out all outcome possibilities simultaneously for children (e.g., showing them all the marbles on a plate) provide opportunity for such analysis. Children typically do well on these.

Another probability task derived probability from sampling – a distribution of variable outcomes over time, which ostensibly requires greater attention to the detail of that distribution, where frequencies of outcomes needed to be observed over many trials. Sequential probability tasks are modeled on this, conforming to the way in which probability is often encountered in everyday life, where we may not have an *a priori* idea of the likelihood of an outcome, or indeed even of the fact that the outcome is variable, until we begin to experience the situation. Even though children do not do so well on these tasks, due to higher processing demands, these skills still link to probability understanding (Bayless and Schlottmann, 2010).

Probability understanding *per se* comes prior to the ability to calculate probabilities, which is largely established in early years (Acredolo et al., 1989; Bryant and Nunes, 2012). Children's understanding of how to quantify it may be restricted to simple relations like "more" or "larger," as Bryant and Nunes showed in their large-scale intervention that more refined proportional reasoning is highly trainable regardless of children's initial ability and that training is effective during the elementary years, indicating that it too is within children's competence in this age range.

A task with lower computational demand -appropriate for the age range- was needed. Therefore, the 'randomness' task was added to the battery to address the fact that sometimes outcomes are determined and predictable, while in other situations they may be unpredictable or potentially random (Reyna and Brainerd, 1994; Bryant and Nunes, 2012). Children seem to make this distinction from ages 4 or 5, as shown by Kuzmak and Gelman (1986), who presented children two devices: one deterministic (marbles lined up in a clear tube, with the first coming out on each trial) and one a lottery device (a cage full of spinning marbles). Children understood that in the first device each outcome is known, but in the second, it is not. Study 1 here employed a similar task, the distribution of target cards in shuffled and unshuffled decks, with an anticipation that this would be sensitive even to the youngest children's abilities.

Altogether, this study included three probability tasks, with different levels of processing complexity. These tasks may elicit variation in performance at different ages and clarify which task might be related to which aspect of thinking about continuous causal processes, such as relative "definiteness" of effect (e.g., stones are very likely to sink, berries and grapes are less likely to) or, as noted earlier, unobservable causal mechanisms (i.e., some other factor affects the relative probability of sinking).

ON THE LINK BETWEEN COVARIATION INFORMATION AND CAUSAL PROCESSES

Grasping bivariate distributions may be more demanding than univariate distributions, because children must track variation in not just one, but two variables, and recognize whether this indicates a link between them. In probability tasks, instead, children need to evaluate the likeness of an event, where the ratio varies between impossibility and certainty. Detection of such links would clearly be helpful in identifying potentially causal variables. For instance, in Schulz et al.'s (2008) study, preschoolers were shown pairs of gears (B and C) operating with a causal chain and a common cause structure on the basis of observing interventions between them. Children as early as 4 years old could discriminate between causal chain and common cause structures (see also Shultz and Mendelson, 1975; Shultz, 1982; Schulz et al., 2008; Sobel et al., 2009).

Considering the Humean regularity and Kantian generative mechanism approaches, Schulz (1982) worked with 3- to 13-year-olds. He reported five experiments, where, for instance, sound, wind, and light transmissions were presented to children in different procedures to assess the essential meaning of causation for children. Children received problems on each of these apparatuses: transmission from source, temporal contiguity versus generative transmission, spatial contiguity versus generative transmission, and covariation. Similar to Ahn et al.'s (1995) findings, he found that children consistently prefer generative mechanism rather than covariation information when they see a conflict between them. For instance, children's justifications were mostly based on mechanism, but rarely based on covariation, even when 3-year-olds' verbal abilities were

TABLE 1 | A typical 2 × 2 contingency table, with cause and effect as the two variables.

	Blossomed	Dead
Plant received fertilizer	AB	A not B
Plant did not receive fertilizer	Not AB	Not A not B

poorer than the elders at the generative aspects of the problem. However, contrary to Ahn et al.'s (1995) proposal, Schultz's results showed that the tendency to analyze causal mechanism is not restricted to prior knowledge – whether children were familiar with the objects or with transmission rules (see also Koslowski et al.'s (1989), for supporting evidence with college students). These studies showed that children can grasp causal relations in the absence of probability or covariation information (see also Perales et al., 2010).

The interest is typically on whether children grasp the implications of covariation information about distinct events for causation. These studies mostly compare the simple case of two potential causes, one regular and one irregular covariate of the effect. To reduce processing demands, only minimal information is given, on whether a cause always co-occurs with the effect (AB cases), or whether in some instances a cause occurs without the effect (A not B cases). If both frequencies are considered, one can derive the probability of the effect, given the cause. This, however, is only part of true covariation assessment, which also requires consideration of the base rate, the conditional probability of the effect occurring in the absence of the cause (i.e., not AB versus not A not B cases) in terms of a 2 × 2 contingency table, as shown in **Table 1**.

The literature focused on covariation (or contingency) judgment therefore considers how humans utilize information from all four cells. A well-established approach is based on the delta *p* statistic (Jenkins and Ward, 1965; Dennis and Ahn, 2001; Marsh and Ahn, 2009), which is the difference between the two probabilities discussed above (the probability of the effect given a cause and the conditional probability of the effect occurring in the absence of the cause). Adult covariation judgment is often studied by providing numerical summaries of the instances in explicit contingency tables, though the instances can, of course, also be presented sequentially, as in the real world, which adds memory demands. To avoid this, and also lower the numerical requirements of such tasks, pictorial formats are typically used with children (see, e.g., Shaklee and Mims, 1981). Note that, as in **Table 1**, these types of studies still illustrate covariation information in causal contexts, to attempt to make complex structured data patterns intuitive and meaningful for children.

Even so, however, children commonly fail to use the delta *p* strategy appropriately, but instead employ simpler strategies that do not consider all four cells of the table or do not weight them evenly. Using this approach, Shaklee and Mims (1981) demonstrated four strategies used by children across development, hierarchically increasing – from the least to the most sophisticated: judgment of the frequency with which the target events co-occur (AB), comparison of the number of times target events do and do not co-occur (AB versus A not B), comparing frequencies of events confirming and disconfirming the relationship (AB plus not A not B versus A not B plus not AB), and optimal assessment of the difference between two conditional probabilities (delta *p*).

These patterns suggest a shift from less to more accurate use of covariation data, where frequency judgment based simply on positive co-occurrence emerges early, while the conditional probability strategy does not appear until the 10th grade. Consistent with this, Shaklee and Paszek (1985) found that, in elementary school, children were most likely to make judgments about covariation by comparing frequencies of the target event and the use of the more advanced strategies identified by Shaklee and Mims (1981) was rare even in fourth grade. Similarly, Ferguson et al.'s (1984) data showed that older children's impressions were influenced more by fuller covariation information rather than frequency information *per se*. In this study, 5- to 13-year-olds were presented with three scenarios about a boy displaying harmful behavior. In condition 1, the harm-doing behavior was low in consistency and also low in frequency. In condition 2, the harm was high in consistency and also high in frequency. In the third condition, the harm was low in consistency but high in frequency. Even preschoolers showed the sensitivity to the frequencies and to the stability of the boy's behavior, but the use of covariation information increased clearly with age.

We hypothesize that primary age children's apparent tendency to focus on frequency over covariation may reflect their difficulties of understanding, but it may also be influenced to some extent by the tasks used. When computational demands, such as ratios and percentages, are minimized, even young children appreciate the difference between variables that co-vary perfectly with an effect or are unrelated to it. For instance, Schulz et al.'s (2008) experimental design with four conditions showed that children can clearly observe a block hitting another block causing it to emit either a train or siren noise. Assessment of imperfect correlation poses more problems, though this is affected too by the way information is presented. For instance, in simple symmetrical tasks (asking whether green or red chewing gum causes bad teeth as illustrated over 10 pictures), even 4-year-olds could evaluate patterns of covariation (Koerber et al., 2005).

To test this hypothesis, we devised a non-causal covariation task to assess whether individual differences in covariation assessment predict children's causal thinking. We kept the task as simple as possible, using a pictorial approach, consistent with the literature, and with our other tasks, we investigated children's assessment of simple covariation patterns. The task included four decks, each consisted eight cards, in which a particular surround shape (a circle or square) contained a particular symbol inside (a star or a moon). Attention focused throughout the degree of co-occurrence between stars and circles. As shown in **Figure 1**, in the first deck, the co-occurrence between stars and circles was 75%. In the second deck, co-occurrence was 50% (analogical to A not B cases). In the third, it was 100% (AB cases). For each deck, children were requested to answer verbally whether a star went together with a circle. Further, they were asked to evaluate how likely a star went with a circle. To answer this question, children were presented with a paper showing a line starting from "never

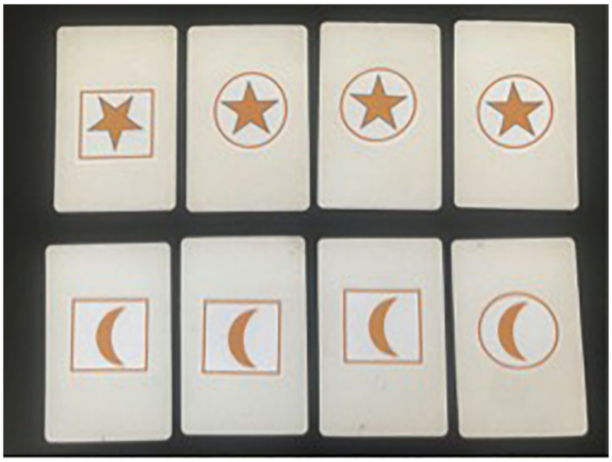


FIGURE 1 | One of the four trials of the covariation task employed in Study 2, displaying the first deck with 75% co-occurrence between stars and circles, representing imperfect covariation.

go together" to "always go together." For each deck, children ticked on this line where they think the likelihood would be best represented. They were encouraged to answer the question by thinking with percentages as well. The focus of this task was on whether children could extract this relation in a number of problems differing in the relative frequency of co-occurrence. We elicited simple verbal and non-verbal responses, but did not ask children to explain their responses, or question them about more complex data.

Overall, in the present study, we employed five probability and covariation tasks, aiming to obtain reliable measures of individual variation in children's statistical thinking to map onto variation in causal inference. Children had to assess frequency relationships *per se*, not frequency relations between cause and effect.

STUDY 1

Study 1 tested the above hypotheses by working with 5- to 11-year-old children. The study employed three causal tasks in relation to continuous processes, one spatial–temporal ability task, one covariation and three probability tasks, and measures of verbal and non-verbal ability as controls.

Methods

Design

The study utilized a combined cross-sectional and individual differences design, employing three groups spanning the English primary (elementary) school age range. We focus here on 10 tasks that were given to children in fixed order within a single one-to-one session: measures of verbal and non-verbal ability, three mini-experiments focusing on causal thinking, and a spatial–temporal task, plus the three probability and covariation tasks.

One-way ANOVAs were used to test for differences between age groups on each task. Fitness of the regression models initially tested by looking at linear, logarithmic, and quadratic trends. Pearson and partial correlations (controlling for age) showed the strength of the associations between the measures. Estimates of the unique variance explained by each predictor task in causal measures were tested using hierarchical linear regressions. Adjusted R^2 values showed the variances explained by the final models. Possible confounds in these estimates were checked with mediation analyses. Combined patterns were tested using path analysis.

Participants

The sample comprised of 107 children, recruited with parental consent from schools in London and Oxford: 35 of them from year 1 (Y1, M_{age} = 6.1 years, sd = 4.4 months), 33 from year 3 (Y3, M_{age} = 8.4 years, sd = 5.9 months), and 39 from year 5 (Y5, M_{age} = 10.3 years, sd = 5.9 months). The sample encompassed wide ethnic and linguistic variation but was skewed toward the upper range in terms of socioeconomic background.

Materials and Procedure

Testing took place out of class in a quiet area within school and, for the tasks described here, lasted on average 35 min per child. Responses were recorded manually on score sheets, but children's replies during the causal tasks were also audio-recorded.

The causal tasks were developed by the authors for this particular project and focused in turn on two contrasting instances of sinking (a stone and a grape sinking), absorption (a piece of tissue and blotting paper absorbing water), and solution (rock and table salt dissolving in water). Comparison between these instances revealed differences, as one item sank slow, another fast, which may then be linked to concurrent differences and commonalities between the objects (e.g., the stone is heavier than the grape, but they are of similar size), which would not be salient in an individual instance.

The tasks were administered and scored as described in Dündar-Coecke et al. (2019). Children were asked to predict outcomes ahead of witnessing simultaneous demonstration of the two instances, which they were then asked to describe, and to explain, as a measure of causal inference assessing the identification of basic factors, operative variables, and mechanisms. Two types of measure were computed from these tasks: totals for accurate prediction from prior knowledge and description for each of the instances considered (maximum = 6) and for inference (ascending score for level of response for each task; maximum = 9); and a total score for causal performance across these indices (alpha = 0.751), which could range from 0 to 21. Interest centers here on the overall causal measure and the measure of inference as the key component where sensitivity to probability and covariation might be anticipated to have an influence.

Appendices 1, 2 provide the full details of task administration and scoring. To confirm reliability, two authors subsequently scored all responses independently from the audio-recordings. Agreement rate was 93%, and final scores were assigned following discussion and checking the audios in the small number of instances where there was a difference. Examples for response levels can be seen in **Supplementary Table 13** in **Supplementary Appendix 2**.

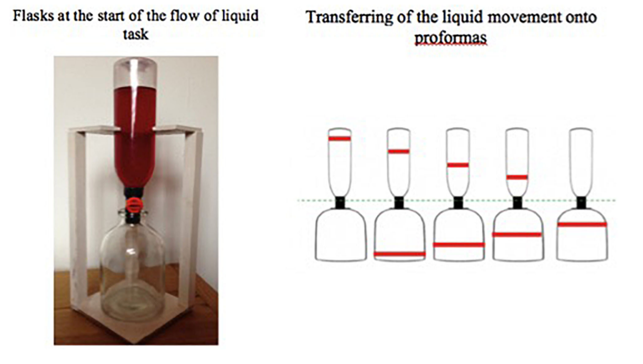

Flasks at the start of the flow of liquid task
Transferring of the liquid movement onto proformas

FIGURE 2 | Flasks at the start of the flow of liquid task.

Measure of spatial–temporal analysis

The *flow of liquid* (FOL) task, adapted from Piaget (1969/2006), examined children's ability to analyze the FOL from one container to another at successive time points and to reconstruct the sequence of change. It consisted of three stages. At the first, two flasks were presented one on top of the other with a tap between (**Figure 2**). The upper flask (I) was filled with red-colored water, while the lower (II) was empty. Children were given a *pro forma* showing both flasks with a space between them, and they marked the respective levels in the flasks by drawing horizontal lines on the *pro forma*. The liquid was then allowed to flow from I to II in four further steps, and the child marked the liquid level on a fresh *pro forma* each time, being invited to correct any errors. At stage two, the five proformas were shuffled and the child put them in order, again being invited to correct any errors. At the third stage, each *pro forma* was cut in two, separating drawings of I from II, shuffled, and the child attempted to put them in order again. Children were expected to match the upper and lower bottles correctly and also put them in the right sequential order. Scores were based on the number of drawings in the correct position at this stage and could therefore range from 0 to 10.

Understanding of probability and covariation

The *randomness* task was used to explore children's understanding of the consequences of a chance mechanism. Participants were shown two identical decks of 30 cards, five of which had smiley face stickers, with the remainder blank. The cards with the stickers were placed at the top of each deck, face up, so that they were visible. One of the decks was then shuffled so that the cards with smiley faces were now mixed with the blank cards. The two decks were then put face down, and participants were asked: "If you want to make sure to pick a smiley face, which deck would you pick from, and why?" Children's choices were marked as 0 or 1 depending on whether they chose the shuffled or unshuffled deck, and if they made the correct choice, their explanations were marked as 0 or 1 according to whether they were able to identify the predictability of the position of the cards with the smiley faces as key to making a choice. Scores could range from 0 to 2.

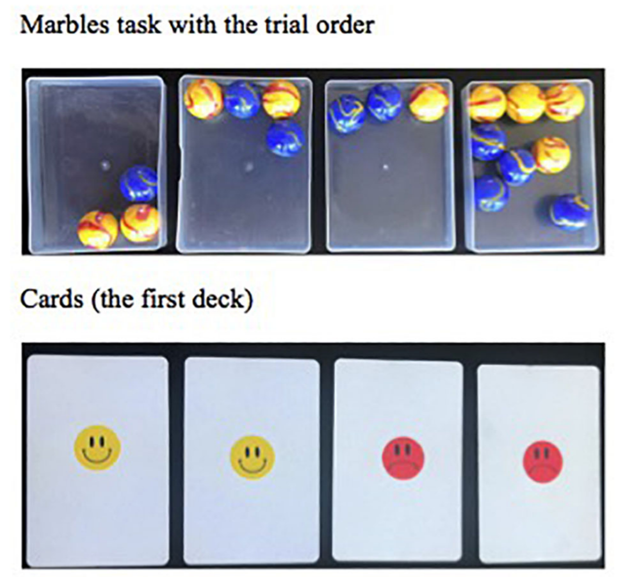


FIGURE 3 | (a) The four trays for the four trials of the Marbles task, in trial order; (b) the deck shown in the first trial of the cards task – only one smiley and one frown were dealt face up, the other two cards were shown face down.

The *marbles task* was adopted from Piaget and Inhelder (1975) to evaluate children's understanding of proportions without sampling. Children were shown over four trials four trays with different numbers of colored marbles (see **Figure 3**). After being told that blue marbles were the winners, children had to say how good each tray was for winning if one marble was picked with eyes closed. They were also asked to estimate how likely they would be to pick a winner from each and could express their answer verbally as either fractions/ratios (as some older children did spontaneously), or by ticking on a line from "never get one" to "always get one." Fully correct answers on both parts of the question were scored as two points for each tray, and partially correct scored as one. Participants who gave consistent correct answers for the second and the fourth tray received an extra two points for confirming verbally that the proportions were identical. This yielded an overall score ranging from 0 to 10.

The *cards task* was developed by the authors to assess children's understanding of frequencies based on sampling. Children saw over four trials four decks comprised of different numbers of cards with smiley versus sad face stickers: (1) two smiley, two sad (see **Figure 3**); (2) two smiley, four sad; (3) four smiley, two sad; and (4) four smiley, four sad, thus utilizing the same proportions as in the marbles task, to ensure that any differences in difficulty between marbles and card tasks did not just reflect differences between samples presented. On each trial, they saw half of the cards dealt out face up, selected to represent the overall proportions, with the others remaining face down. Children had to say how good each deck was for picking a smiley, and then like the marbles, estimate the chances of doing so. Scores were similar as to the marbles task and could range from 0 to 10.

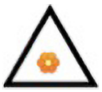


FIGURE 4 | Practice display of the covariation task.

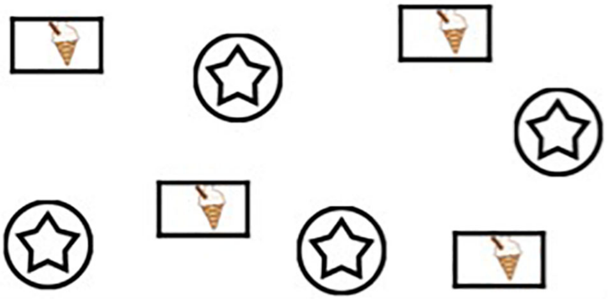


FIGURE 5 | First display of covariation task representing perfect covariation.

The *covariation task* was developed by the authors. The task involved three trials on a laptop, each displaying, in pseudorandom order, a series of eight pictures, half in a square and half in a circle frame. Children had to detect whether there was a relation between frame shape and content of the picture.

The task was started with an introduction, displaying a triangle together with a flower (**Figure 4**). Children were told "Now, you are going to see some shapes appear one by one on the screen filled with different pictures like this following: a flower goes with a triangle. I will ask you each time look at the screen carefully and tell me what shape goes with what picture."

The first display showed perfect covariation: four pictures of an ice cream in the square and four of a star in the circle (see **Figure 5**). Each figure appeared on the screen one by one, and children were asked: *Which shape goes with the ice cream? Do they always go together? Which shape goes with the star? Do they always go together?*

The second display showed imperfect covariation (75% contingency): three pictures of a basketball and one of sunglasses in the squares, and three of a phone and one of a line in the circle. Participants were asked: *Which picture goes with the circle? Do they always go together? Which picture goes with the square? Do they always go together?* The third display had no pattern (zero contingency), the circles and squares all contained different pictures, and participants were again asked the same questions. Shapes were kept consistent to provide a common anchor across displays, but pictures were varied, to avoid carry-over. All trials consisted of eight figures each. Co-occurrence could be expressed as fractions/ratios, or by ticking on a line, as for marbles, from "can't tell at all" to "definitely." Each correct answer was marked as 1 point. Children were expected to identify of the dominant correlate for displays 1 and 2, and they were supposed to say "none"/"any" for display 3 based on the appropriate estimation of the strength of association. Scores could range from 0 to 12.

Measures of verbal and non-verbal ability

The expressive vocabulary and block design subtests from the Wechsler Abbreviated Scale of Intelligence (WASI) (Wechsler, 2011) were used to provide standard measures of verbal and non-verbal ability.

The WASI vocabulary is a measure of expressive language, word knowledge, and verbal concept formation. Children were required to define the words when the researcher read aloud. Administration and scoring followed standard procedures.

The WASI Block Design is a subset to explore children's non-verbal cognitive abilities. Children were shown nine red and white square blocks and a book illustrating different patterns in each page that could be made with the blocks. Children were asked to arrange the blocks to match each design shown in the picture, increasing in difficulty. This task aimed to measure children's ability to analyze and synthesize abstract representations within specific time limits for each display.

Results

Analyses utilized data from the 107 participants who completed testing, except where noted. Age trends on each measure are presented below, followed by analyses of relationships between the causal, spatial–temporal, probability, and covariation measures. All statistical tests were two-sided where the highest p value was set to 0.05. Employing the F and t test procedures based on the general linear model of regression, the observed power for the regression was 0.95, which was calculated using G*Power 3.1.9.2 (Erdfelder et al., 2007).

Developmental Trajectories

There were a significant negative skew on total causal score, FOL, randomness, marbles, and cards and a positive skew on block design, due to the youngest and oldest age groups, respectively, exhibiting a longer tail of scores; inference, covariation, and vocabulary were normally distributed. **Figure 6** demonstrates the developmental trajectory for each measure using scores standardized to a scale between 0 and 1 for comparability. Overall, there was a clear upward trend for all tasks, but with variation in relative difficulty. Block design was in particular difficult for most children. Significant negative skew indicated that most children failed to gain higher scores in this task, while FOL was easier (hence the difference in direction of skew), with causal total lying in between. Comparing the trends, the steepest gradients were seen for blocks and covariation, followed by randomness, cards, and marbles.

Means and standard deviations on the original scales are shown in the **Table 2** (study 1). In terms of differences between age groups, marbles, FOL, causal total, and vocabulary tasks showed a similar pattern: the steepest increase was between Y1 and Y3, with a slow down subsequently. For the rest, the steepest gradient was between Y3 and Y5, except randomness where growth was linear. One-way ANOVAs by school year found highly significant increases with age on all variables, using the Welch robust statistic, $p < 0.001$ in each case. The majority of inference responses on all three causal tasks focused solely on relevant factors or variables (scores of 1 or 2), though mechanism responses were more evident among older children: 2.9% of

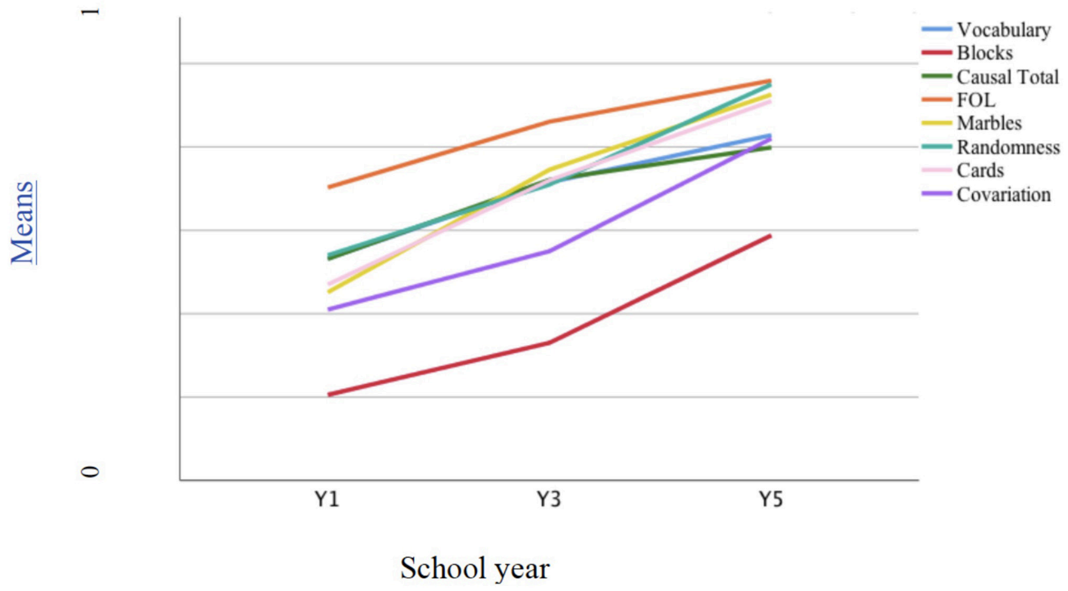


FIGURE 6 | Developmental trajectories of all variables computed using standardized measures across the three year groups.

children in Y1 gave one or more mechanism response, 24.2% in Y3, and 30.8% in Y5 (cf. Dündar-Coecke et al., 2019, for the response profiles of the causal tasks).

Relationships Between Causal Performance and Spatial–Temporal Analysis, Probability, and Covariation

Correlations between variables

The relationship of the predictor variables to the causal measures was linear, apart from block design, where it was logarithmic (R^2 for linear fit = 0.263; while it was 0.368 for logarithmic trend). Zero-order Pearson correlations between the different measures showed overall causal performance and inference was strongly positively associated with all the potential predictors, which were themselves all positively correlated with each other (**Table 3**, study 1 correlations). The high correlations between causal total and inference were plausible, as causal total contained prediction, description, and inference scores across the three tasks.

When age in months and verbal (vocabulary) and non-verbal ability (log block design) were controlled for, only FOL, marbles, cards, and covariation remained significantly associated with total causal performance. The same set of variables was also related to inference, with the exception of cards. FOL was related to both cards and marbles and to covariation to a lesser extent. The probability and covariation measures were predominantly related to each other, though marbles and cards were the most closely related measures, with covariation – and randomness – more distinct from these. Randomness had little relation to the causal measures, possibly because its narrow scoring range made it less discriminating. It did not affect the beta values of other variables and remained non-significant in each regression model and was therefore discounted from further consideration.

Hierarchical regression models

Hierarchical regression was used to examine the unique variance accounted for by the remaining predictors. Taking total causal score and inference in turn as the dependent variable, age in months and vocabulary were entered in the first stage of the analysis. Marbles, cards, and covariation were entered after the control variables, but with marbles first, since it related best to the causal indices; this made it possible to assess its specific impact before including cards and covariation. Log block design was entered at the fourth stage, in order to assess the influence of verbal and non-verbal ability separately and to examine the predictive power of the statistical measures before and after non-verbal ability was controlled for. The spatial–temporal measure, FOL, was entered at the fifth stage, since it appeared to be the most robust predictor overall. Analyses for prior knowledge and description with the same order of entering predictors are presented in Appendix 5.

For *total causal score*, the analysis (**Table 4**, study 1 regressions) produced significant ΔR^2 at each stage except the third. Age and vocabulary were significant predictors at the first stage, but the beta for vocabulary dropped and age was superseded by marbles when that was entered. Vocabulary and marbles dropped out when cards and covariation were added, but neither of the latter was significant, indicating that all four predictors shared variance. The beta for cards was smaller than that for covariation, which was marginally the largest remaining predictor. Log block design was a significant predictor when added at the fourth stage, and produced further drops in the betas for all the other variables, with a bigger impact on covariation than marbles or cards. FOL joined log block design as a further predictor at the final stage, without substantially affecting the betas for the other variables, except marbles.

TABLE 2 | Study 1 mean scores (with standard deviations) on total causal performance (max = 21), inference (max = 9), vocabulary (max = 43), block design (max = 58), flow of liquid (FOL; max = 10), randomness (max = 2), marbles, cards (max = 10), and covariation (max = 12).

	Y1	Y3	Y5	Total
Causal total	10.63 (4.44)	14.42 (2.96)	15.97 (2.44)	13.75 (4.04)
Inference	3.34 (1.89)	5.06 (1.54)	5.54 (1.54)	4.67 (1.90)
Vocabulary	22.89 (5.29)	30.76 (5.86)	35.62 (5.20)	29.95 (7.59)
Blocks	11.91 (6.08)	19.15 (9.52)	34.10 (13.25)	22.23 (13.86)
FOL	7.03 (3.27)	8.61 (2.09)	9.59 (1.31)	8.45 (2.55)
Randomness	1.09 (0.89)	1.42 (0.79)	1.90 (0.38)	1.49 (0.78)
Marbles	4.51 (3.08)	7.45 (2.95)	9.26 (1.82)	7.15 (3.29)
Cards	4.69 (3.56)	7.15 (2.76)	9.10 (1.59)	7.06 (3.27)
Covariation	4.83 (2.18)	6.61 (3.29)	9.82 (2.81)	7.20 (3.48)

Study 2 mean scores (standard deviation) on total causal performance (max = 33), inference (max = 12), flow of liquid (FOL, max = 12), DTV (max = 18), marbles (max = 10), covariation (max = 8), block design (max = 45), and vocabulary (max = 43).

	Y1	Y3	Y5	Total
Causal total	15.33 (4.76)	16.20 (4.47)	18.72 (2.95)	16.82 (4.31)
Inference	5.03 (2.36)	5.69 (2.56)	6.63 (1.75)	5.82 (2.32)
Vocabulary	22.48 (5.39)	29.05 (5.01)	34.01 (4.56)	28.86 (6.77)
Blocks	12.43 (5.62)	16.45 (7.12)	24.01 (8.97)	17.91 (8.78)
FOL	8.33 (4.34)	9.24 (3.79)	11.16 (2.40)	9.65 (3.72)
DTV	12.58 (3.50)	13.33 (3.08)	14.56 (2.77)	13.54 (3.18)
Marbles	3.76 (2.99)	5.39 (3.05)	7.41 (2.95)	5.61 (3.31)
Covariation	4.48 (1.96)	5.34 (2.35)	6.34 (1.96)	5.44 (2.22)

TABLE 3 | Study 1 zero-order and partial correlations between measures (zero-order correlations above diagonal, $N = 107$; partial correlations below diagonal, controlling for age in months, verbal and non-verbal ability, $N = 106$ due to missing date of birth data for one participant; significant values in bold, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$).

	Causal total	Inference	Vocabulary	Log blocks	FOL	Randomness	Marbles	Cards	Covariation
Causal total	1	**0.90*** **	**0.54*** **	**0.61*** **	**0.52*** **	**0.39*** **	**0.55*** **	**0.52*** **	**0.56*** **
Prior	**0.69*** **	**0.53*** **	**0.47*** **	**0.56*** **	**0.46*** **	**0.27** **	**0.42*** **	**0.42*** **	**0.47*** **
Description	**0.79*** **	**0.70*** **	**0.44*** **	**0.48*** **	**0.45*** **	**0.40*** **	**0.42*** **	**0.49*** **	**0.43*** **
Inference	**0.85*** **	1	**0.47*** **	**0.52*** **	**0.42*** **	**0.34*** **	**0.53*** **	**0.43*** **	**0.51*** **
Vocabulary	–	–	1	**0.68*** **	**0.44*** **	**0.44*** **	**0.52*** **	**0.53*** **	**0.64*** **
Log blocks	–	–	–	1	**0.43*** **	**0.41*** **	**0.56*** **	**0.54*** **	**0.62*** **
FOL	**0.30** **	**0.20* **	–	–	1	**0.35*** **	**0.55*** **	**0.56*** **	**0.47*** **
Randomness	0.13	0.10	–	–	0.16	1	**0.49*** **	**0.54*** **	**0.43*** **
Marbles	**0.25* **	**0.28** **	–	–	**0.38*** **	**0.28** **	1	**0.76*** **	**0.63*** **
Cards	**0.21* **	0.13	–	–	**0.38*** **	**0.36*** **	**0.61*** **	1	**0.60*** **
Covariation	**0.20* **	**0.21* **	–	–	**0.20* **	0.15	**0.37*** **	**0.32** **	1

Study 2 zero-order and partial correlations between measures (zero-order correlations above diagonal, partial correlations below diagonal, controlling for age in months, verbal and non-verbal ability, $N = 124$; significant values in bold, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$).

	Causal total	Inference	Vocabulary	Log blocks	expFOL	DTV	Marbles	Covariation
Causal total	1	**0.89*** **	**0.53*** **	**0.48*** **	**0.60*** **	**0.44*** **	**0.39*** **	**0.43*** **
Inference	**0.82*** **	1	**0.55*** **	**0.49*** **	**0.61*** **	**0.49*** **	**0.42*** **	**0.41*** **
Vocabulary	**–**	–	1	**0.53*** **	**0.49*** **	**0.39*** **	**0.48*** **	**0.43*** **
Log blocks	**–**	–	0.53**	1	**0.50*** **	**0.42*** **	**0.55*** **	**0.34*** **
expFOL	**0.41*** **	**0.41*** **	–	–	1	**0.54*** **	**0.52*** **	**0.43*** **
DTV	**0.23* **	**0.30** **	–	–	**0.38*** **	1	**0.44*** **	**0.39*** **
Marbles	0.10	0.14	–	–	**0.29** **	**0.23* **	1	**0.40*** **
Covariation	**0.24** **	**0.20* **	–	–	**0.25** **	**0.25** **	**0.20* **	1

TABLE 4 | Study 1 hierarchical regression analysis with *total causal* score as dependent variable (significant predictors in bold).

Model	M1	M2	M3	M4	M5
Predictor		**B**			
Age in months	**0.332****	0.207	0.176	0.115	0.109
WASI vocabulary	**0.310****	**0.231***	0.154	0.059	0.044
Marbles		**0.310****	0.172	0.145	0.096
Cards			0.104	0.086	0.033
Covariation			0.184	0.131	0.121
Log blocks				**0.284***	**0.273***
Flow of liquid					**0.203***

AdjR^2 = 0.454; ΔR^2 = 0.347*** for M1; 0.061** for M2; 0.022 for M3; 0.035* for M4; and 0.026* for M5. *p < 0.05.**p < 0.01.***p < 0.001.

Study 2 hierarchical regression analysis with total causal score as dependent variable (significant predictors in bold).

Model	M1	M2	M3	M4	M5	M6
Predictor		**β**				
Age in months	–0.136	–0.171	–0.174	–**0.249***	–**0.230***	–**0.218***
WASI vocabulary	**0.624*****	**0.551*****	**0.483*****	**0.434*****	**0.408*****	**0.349****
Marbles		**0.202***	0.147	0.050	0.017	–0.053
Covariation			**0.221****	**0.209***	**0.176***	0.130
Log blocks				**0.284****	**0.250***	**0.197***
DTV total					0.160	0.062
Expflow of liquid						**0.349*****

*AdjR^2 = 0.459; ΔR^2 = 0.292*** for M1; 0.031* for M2; 0.037** for M3; 0.046** for M4; 0.018 for M5; and 0.066*** for M6. *p < 0.05.**p < 0.01.***p < 0.001.*

The analysis for *inference* produced similar outcomes at the first two stages (**Table 5**, study 1 regression), except age and vocabulary that were both superseded by marbles. In this case, however, the addition of cards and covariation had little appreciable impact on marbles. Covariation had the second largest beta, but was not significant. The inclusion of log block design had little impact on marbles, cards, and covariation. The addition of FOL had somewhat more impact on marbles, but the latter remained the sole significant predictor.

Overall, the regression analyses revealed clear overlaps between the influence of all the predictors on causal performance. However, the relative impact of including FOL and log block design in the models indicates marbles and cards were somewhat more closely related to the former and covariation to the latter. Probability and covariation therefore appeared to capture somewhat different dimensions, in line with the partial correlations. In particular, while spatial–temporal and non-verbal ability were the strongest predictors of overall causal thinking, for inference, the effects of probability were stronger.

Nature of shared variances between predictors

Factor analysis with varimax rotation was used to explore in more depth the nature of the relationship between FOL, log block design, marbles, cards, and covariation, given their shared influence on the causal indices. The Kaiser–Meyer–Olkin (KMO = 0.834) measure of sampling adequacy was well within acceptable limits. The KMO identified a four-factor solution that explained 95% of the shared variance between the five measures, which confirmed separable components relating to marbles/cards, covariation, FOL, and log block design (**Table 6**, study 1 rotated component matrix).

In view of this, maximum likelihood path analysis was used to examine whether there were specific directional relationships between the predictors that would explain the observed patterns of overlap in their influence on the causal indices. For both causal measures, the best fit was provided by an extended mediation model, which was assessed by the chi-squared and probability values penalized by the Akaike information criterion (AIC), where the fitness of the model improved as the AIC value lowered. The best fit was χ^2 = 3.891, p = 0.273 for total causal and χ^2 = 3.887, p = 0.274 for inference, with df = 3 for both. **Figure 7** illustrates the model and path coefficients obtained for total causal score. Black and gray paths were used to distinguish between subsidiary and major path coefficients. The model illustrated a stable pattern of effects in which non-verbal ability, awareness of covariation, and probability (as indexed by marbles) support spatial–temporal analysis, but with each also influencing aspects of causal reasoning to different degrees. For overall causal performance, non-verbal and spatial–temporal ability have the largest direct effects, with the effects of probability and covariation smaller by comparison; for inference, the direct effect of probability, 0.219, is stronger than non-verbal and spatial–temporal ability, 0.190 and 0.101, respectively. Age and vocabulary have little or no direct

TABLE 5 | Study 1 hierarchical regression analysis with *inference* as dependent variable (significant predictors in bold).

Model	M1	M2	M3	M4	M5
Predictor		β			
Age in months	**0.285***	0.144	0.127	0.083	0.080
WASI vocabulary	**0.272***	0.182	0.126	0.060	0.050
Marbles		**0.351****	**0.321***	**0.301***	**0.273***
Cards			−0.068	−0.081	−0.111
Covariation			0.190	0.153	0.147
Log blocks				0.201	0.194
Flow of liquid					0.118

AdjR^2 = 0.339; ΔR^2 = 0.262*** for M1; 0.078** for M2; 0.017 for M3; 0.018 for M4; and 0.009 for M5. *p < 0.05.**p < 0.01.***p < 0.001.

Study 2 hierarchical regression analysis with *inference* as dependent variable (significant predictors in bold).

Model	M1	M2	M3	M4	M5	M6
Predictor		β				
Age in months	−0.174	−**0.215***	−**0.217***	−**0.291****	−**0.265****	−**0.253****
WASI vocabulary	**0.667*****	**0.583*****	**0.531*****	**0.482*****	**0.445*****	**0.392*****
Marbles		**0.237****	**0.195***	0.100	0.053	−0.010
Covariation			**0.169***	0.157	0.110	0.069
Block design (log)				**0.280****	**0.233***	**0.184***
DTV total					**0.224****	0.135
Expflow of liquid						**0.317*****

*AdjR^2 = 0.488; ΔR^2 = 0.318*** for M1; 0.042** for M2; 0.022* for M3; 0.044** for M4; 0.035** for M5; and 0.054***for M6. *p < 0.05.**p < 0.01.***p < 0.001.*

TABLE 6 | Study 1 four-factor model for flow of liquid, log blocks, marbles, cards, and covariation (significant predictors in bold).

	Factor 1	Factor 2	Factor 3	Factor 4
FOL	0.292	**0.927**	0.163	0.168
Log blocks	0.281	0.172	**0.907**	0.264
Marbles	**0.821**	0.233	0.234	0.298
Cards	**0.856**	0.257	0.223	0.208
Covariation	0.335	0.189	0.287	**0.876**

Study 2 three factor solution for exponential flow of liquid, DTV, log block design, marbles, and covariation.

	Factor 1	Factor 2	Factor 3
Exp FOL	0.461	**0.648**	0.252
DTV total	0.191	**0.915**	0.152
Log blocks	**0.857**	0.228	0.080
Marbles	**0.794**	0.233	0.246
Covariation	0.194	0.213	**0.951**

impact on causal thinking in these models and act as background variables, influencing the main predictors to different degrees.

Further moderation analyses confirmed there were no interaction effects between log block design, marbles, covariation, or FOL in predicting causal scores.

Discussion

This study confirmed developmental trends in the ability to analyze probability, covariation, and spatial–temporal information, with clear increases across the age groups, though the statistical indices used here showed later growth than the spatial–temporal indices, with children approaching ceiling by Y5 on FOL.

Performance on FOL remained discriminating, but the use of statistical information consistently correlated with both overall causal thinking and inferential level causal analysis. Factor analysis confirmed these are distinct competences, though classical and frequentist probability was found to be closely related. Probability, covariation, and spatial–temporal analysis had, in part, independent effects on causal inference, but also, in part, interrelated influence connected to non-verbal ability.

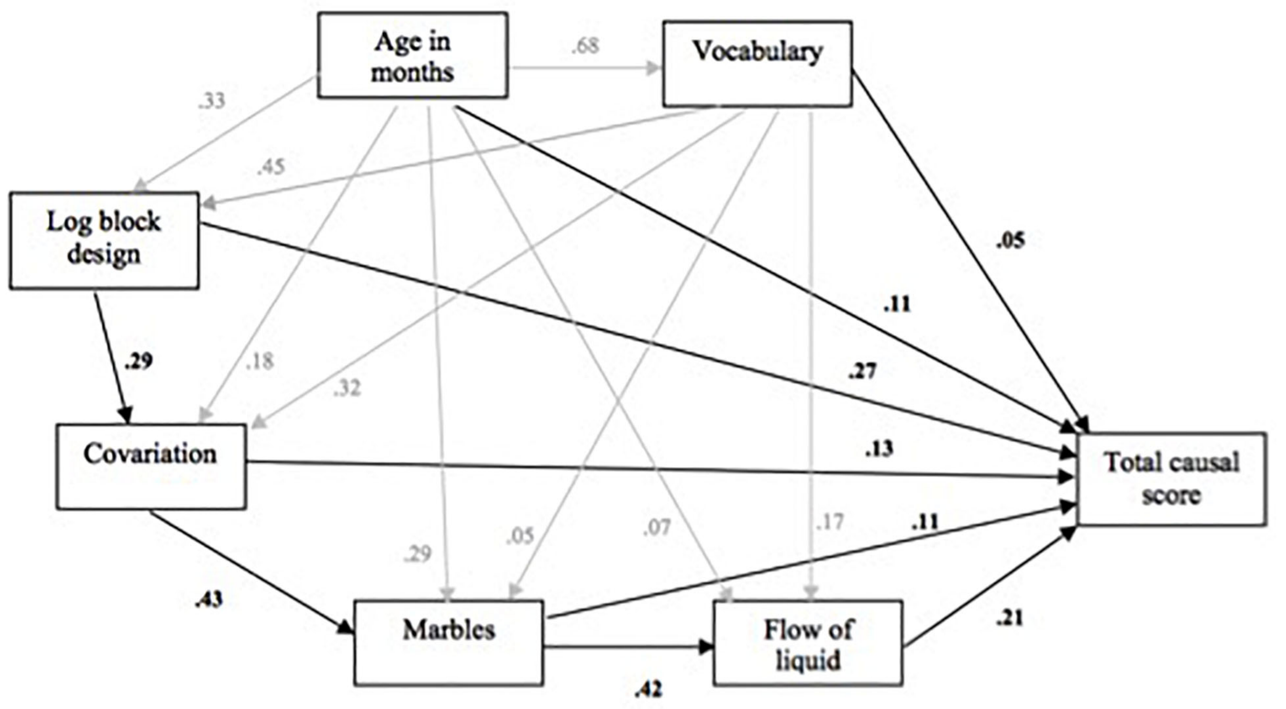


FIGURE 7 | Extended mediation model for the effects of log block design, covariation, marbles and flow of liquid on total causal score (subsidiary effects in gray).

Marbles performance was a significant predictor to begin with in both regression analyses, but for overall causal thinking, it dropped substantially with the inclusion of cards and covariation, and then again when FOL was added. For inference, however, it remained a significant predictor once included, though it was again affected by FOL. Covariation was never a significant predictor. This might be partly attributable to the limited number of steps involved in the task we used affecting its sensitivity – but it nevertheless had a sizeable beta until log block design was included, and it interacted with FOL at lower levels. Although verbal ability had no impact on any aspect of causal performance, this may have been due to the relatively narrow social range of the sample; there were nevertheless clear indications that statistical ability in particular overlapped in part with verbal ability.

The findings suggested that only spatial–temporal analysis and non-verbal form of cognitive ability significantly associated with causal thinking; neither statistical inference nor verbal ability had significant explanatory power. However, we need to consider the sample and task characteristics before arriving at conclusions. Thus, the next study refined the task battery and examined the replicability of findings among a wide range of population.

STUDY 2

The modified causal tasks followed the structure of a scientific investigation, the FOL task was extended, an additional spatial–temporal measure was derived from an adaptation of Wilkening's (1981) distance/time/velocity integration tasks, and a more socially representative sample within the same age range was employed.

The marbles task remained to assess children's probability judgments. Another covariation task, which was a revision of that used in study 1, utilized physical materials in the form of decks of cards rather than a computer display, in keeping with most of the other test materials. The tasks were selected on the basis of their relative predictive strength in study 1. Therefore, the cards task was dropped, in view of its overlap with marbles in study 1, and randomness was dropped because of its lack of predictive power.

Methods

Design

The design, age groups, task order, and administration were all equivalent to study 1. We focus on nine tasks given in fixed order within a single one-to-one session: WASI expressive vocabulary and block design (Wechsler, 2011), three causal experiments, and two spatial–temporal tasks, plus the probability and covariation tasks at the end.

Participants

The sample comprised 124 children, recruited with parental consent from three schools in Oxford: 36 from Y1, mean age = 5 years, 11 months, sd = 3.8 months; 45 from Y3, mean age = 7 years, 11 months, sd = 3.6 months; and 43 from Y5, mean age = 9 years, 9 months, sd = 5.1 months. Children's ethnic and linguistic background was similar to study 1 but covered a more broadly representative range of socioeconomic backgrounds.

Materials and Procedure

Testing for the tasks took an average of approximately 42 min per child (min = 29, max = 57). Responses were recorded in the same way as for study 1.

The *scientific method causal tasks* were developed by the authors and administered and scored as described for study 2 in Dündar-Coecke et al. (2020). The tasks followed a more realistic scientific procedure, with a sequence of observation, description, prediction, justification, and explanation. Full details of this protocol can be found in Appendix 3. Briefly, the tasks again focused in turn on contrasting instances of sinking, absorption, and solution, but in this case, children first observed and described two instances before being asked to predict. They then justified their predictions by judging the outcomes of further three items. Children then explained the influences at work across all five instances. Two types of measure were computed from these tasks: totals of each task for accurate description (maximum = 3), prediction and justification (each maximum = 9), and level of explanation (again, assessing identification of basic factors, operative variables, relationships between variables, and mechanisms; maximum = 12). The second measure was the total score for causal performance across all indices, 0–33, alpha = 0.724. Interest again centered on the overall score for causal performance and that for inference.

Appendices 3, 4 provide the details of the scripts and scoring systems. As in the first study, children's responses were scored independently by two authors based on the criteria shown in **Supplemenatry Tables 14, 15** in **Supplementary Appendix 4**. The independent scores were compared for interrater reliability. Any difference in the independent scores was followed by further checking of the audio records, with a discussion to get a 100% agreement on the final scores.

The measures of *verbal and non-verbal ability* and the FOL were all similar as described for study 1, except that six stages were employed for FOL rather than five, and scores could therefore range from 0 to 12. The distance–time–velocity (DTV) measure required children to make estimates of each of distance, time, and velocity in turn, by integrating information about the other two variables. Each task utilized scenarios akin to those employed by Wilkening (1981), displayed on PowerPoint slides. For *distance*, children judged how far three animals varying in speed (cat, mouse, and turtle) would run in a fixed time, counted out by the experimenter, to escape from a barking dog. For *time*, they had to estimate, by counting themselves, how long an animal (cat, bunny, and turtle) would take to run to a fixed point, with the second half of the run concealed behind a wall. For *velocity*, they had to judge which of seven animals (deer, horse, cat, bunny, mouse turtle, and snail) would make it to a fixed destination in a given period of time, counted out by the experimenter. Children's judgments relied entirely on mental projection based on information provided, and no actual motion was observed to support the key elements of these. Each task consisted of three trials, with responses on each trial scored 0–2 in terms of degree of accuracy. The total score across the three tasks could therefore range from 0 to 18.

Measure of probability

The *marbles task* followed exactly the same procedure for administration and scoring as in study 1.

Measure of covariation

The *covariation task* was developed by the authors to provide an alternative approach to the previous computer-based covariation task. The task utilized cards showing one of four images: a circle containing a star, a circle containing a crescent moon, a square with a star, and a square with a moon. Attention focused throughout on the degree of co-occurrence between stars and circles, with the squares as distractors. Children saw four decks in turn consisting of eight cards. In the first deck (**Figure 1**), three of the circles contained stars and one a moon; co-occurrence between stars and circles was therefore 75%. In the second deck, co-occurrence was 50%: half of the circles contained stars and half contained moons. In the third, it was 100%: all circles contained stars. In the fourth, it was 0%: all the circles contained moons.

In each case, the cards were laid out face up before the child in random order, and they were asked to say from what they saw in front of them how often stars and circles went together. As in the marbles task, they made a verbal judgment first of all (e.g., "three times," "always"), and then provided an estimate of frequency by ticking on a line, one end marked "never" and the other "always." Correct answers on both responses were scored as two points for each deck, allowing for some lack of exact precision in the tick responses. Partially correct responses were scored as one. Scores therefore varied between 0 and 8.

Results

Analyses utilized data from all 124 participants. Age trends are presented first, followed by analyses of relationships between causal, spatial–temporal, probability, and covariation measures. All statistical tests were two-sided. The highest p value was set to 0.05. Using the G*Power 3.1.9.2 (Erdfelder et al., 2007), the observed power for the regression was 0.97.

Age Profiles

Mean scores on each measure are shown in **Figure 8**, using standardized measures, as in study 1. There were a significant negative skew on FOL, DTV, and vocabulary and a positive skew on block design; the remaining variables were normally distributed. Again, the developmental trend was clear, with tasks varying in difficulty. Blocks task was in particular difficult for children, while FOL was easier. The causal task was also more difficult than in study 1.

Comparing the increase in scores across year groups, the steepest gradients were apparent for blocks, FOL, and causal total, with the greatest growth between Y3 and Y5 (see **Table 2**, study 2 mean scores). The remaining measures showed a linear trend with marbles exhibiting the steepest gradient. One-way ANOVAs by school year found significant increases with age on all variables, using the Welch robust statistic, $p < 0.01$ in each case, except DTV, $p < 0.05$. For all measures, there were significant increases in scores from Y1 to Y5. Performance on FOL was similar to study 1, again approaching ceiling by Y5,

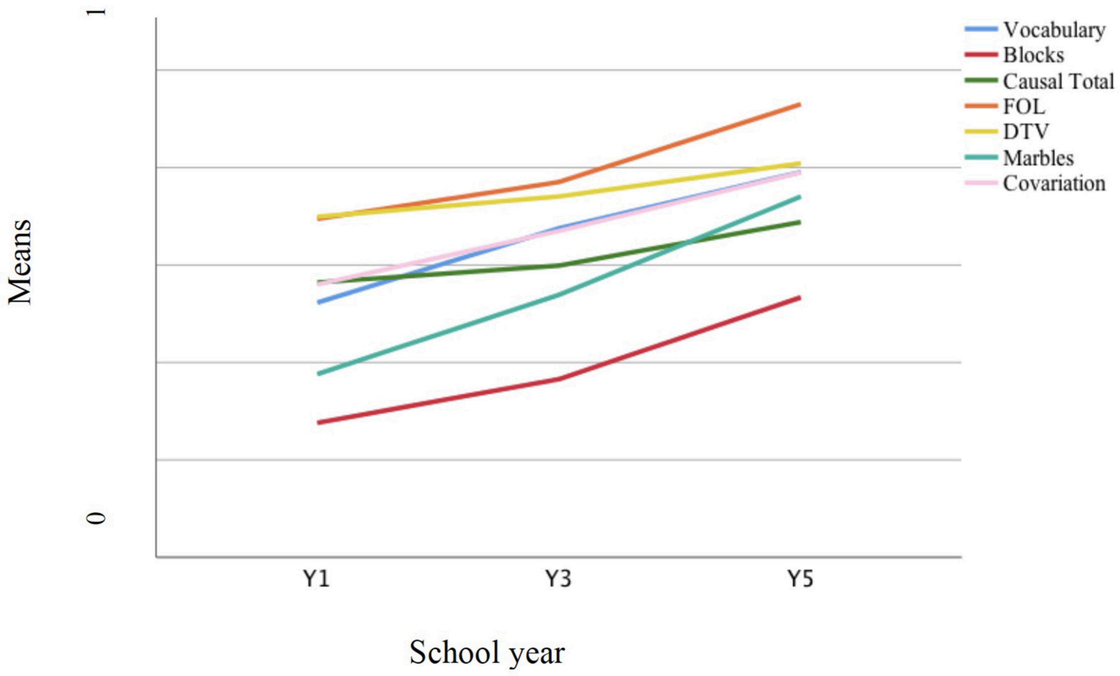


FIGURE 8 | Developmental trajectories of all variables computed using standardized measures across the three year groups.

despite the extra step in the procedure; the later growth on this and the causal indices more probably reflected the mixed sample of study 2 lagging behind the higher socioeconomic status (SES) sample of study 1. In line with this, these children exhibited marginally lower mean scores for vocabulary (albeit with lower variance) and notably lower scores on block design, marbles, and – even allowing for the change in task – covariation. Mechanism responses were also less common – 4.6% of responses in Y1 were at this level, but only 11.1% in Y3, and 17.8% in Y5.

Relationships Between Causal Performance and Spatial–Temporal Analysis, Probability, and Covariation

Correlations between variables

The fitness tests assessing the trends between the predictor and causal variables showed that block design was again logarithmically related to the causal measures [R^2 for linear fit = 0.194 ($F(1, 122) = 29.375$, $p = 0.000$); R^2 for logarithmic fit = 0.232 ($F(1, 122) = 36.849$, $p = 0.000$], and FOL was marginally exponential [R^2 for linear fit = 0.342 ($F(1, 122) = 63.546$, $p = 0.000$); R^2 for exponential fit = 0.357 ($F(1, 122) = 67.810$, $p = 0.000$]; relationships for the other predictors were linear. As in study 1, zero-order Pearson correlations showed the causal indices were strongly positively associated with all the potential predictors, and the predictors were themselves all positively correlated with each other, as shown in **Table 3**, study 2 correlations.

Controlling for age in months, vocabulary, and log block design, only FOL, DTV, and covariation remained significantly associated with total causal score and inference. In contrast to study 1, marbles was unrelated to either causal measure. Marbles and covariation were more weakly related than in study 1, possibly reflecting the revised measure of the latter and the lower SES sample, and there was a more equivalent relationship between each and the spatial–temporal measures; the strongest relationship was between the spatial–temporal measures, FOL, and DTV.

Hierarchical regression models

The predictors were entered into regression analyses for both of the causal indices in equivalent order to study 1; the additional spatial–temporal measure, DTV, was entered in an extra step ahead of FOL. Analyses for description, prediction, and justification with the same order of entering predictors are presented in Appendix 5.

For *total causal score*, the analysis produced significant ΔR^2 at each stage except for the fifth (**Table 4**, study 2 regressions). Vocabulary was a significant predictor throughout, with a substantially higher beta than in study 1, although this dropped at each successive stage. Marbles was a significant predictor when it was entered at the second stage, but it dropped out when covariation was added. Covariation was significant and remained so until the inclusion of exponential FOL. Log block design was a significant predictor when added at the fourth stage, and produced a further drop in the beta of marbles. Age became a significant negative predictor at this stage, possibly due to influence of residual variance. When DTV was included, this led to drops in the betas for covariation and log block design, but it was not significant itself. Exponential FOL joined vocabulary and log block design as a positive predictor at the final stage, and the betas for the other predictors dropped further. This suggested that two cognitive ability measures (vocabulary and block design)

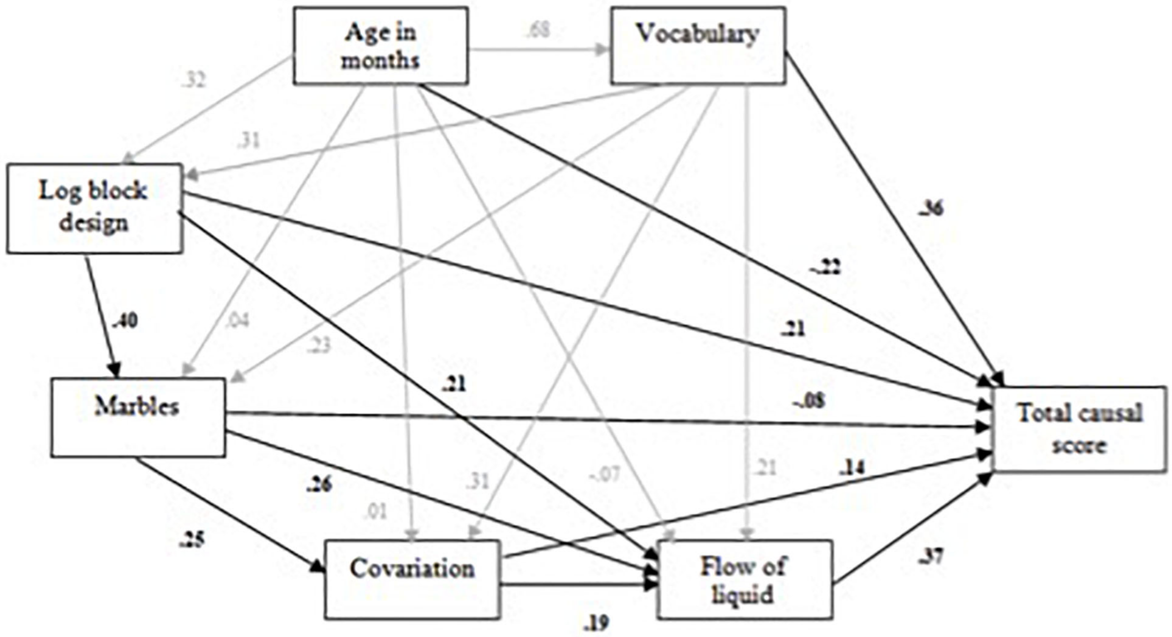


FIGURE 9 | Further extended mediation model for the effects of log block design, covariation, marbles and flow of liquid on overall causal performance (subsidiary effects in gray).

and one spatial–temporal measure played a significant role in predicting total causal score.

The analysis for *inference* (**Table 5**, study 2 regression) produced similar outcomes to that for total causal score, except that marbles stayed significant when covariation was entered at the third stage, as in study 1, albeit with a lower beta than there. Both dropped out with the inclusion of log block design. The addition of DTV reduced the beta for all three of these variables, and it was itself significant. FOL was again a significant positive predictor, and its inclusion reduced the betas for the other variables, most notably DTV, which became non-significant.

Overall, the regression analyses revealed clear overlaps between the influence of all the predictors on causal reasoning. In contrast to study 1, covariation was the strongest of the two statistical measures for total causal performance, but marbles nevertheless remained stronger for inference. As in study 1, there were significant differences in marbles score between children giving differing levels of inference response for sinking, $F = 4.728$, $p = 0.001$; absorption, $F = 2.701$, $p = 0.034$; and solution, $F = 5.371$, $p = 0.001$, with $df = 4, 119$ for each, with effects again restricted to differences between those with lower (0, 1, or 2 here) and higher (3 or 4) inference scores. In this study, however, marbles appeared to be more related to non-verbal ability than to the spatial–temporal measures, while covariation was closer to the latter – as if marbles and covariation had swapped status. Neither of the statistical measures survived to the final models in study 2, and their impact was more noticeable here for overall causal performance than for inference.

In contrast to study 1, verbal ability was consistently a strong predictor, alongside non-verbal and FOL in the final model. The impact of the other spatial–temporal measure, DTV, was relatively modest, with FOL substantially reducing the beta of DTV in both analyses. Further regression analysis with all variables confirmed that only FOL was a significant predictor of total causal score ($\beta = 0.356$, $p = 0.000$) when both cognitive ability measures – vocabulary and block design – were controlled for.

Nature of shared variances between predictors

Factor analysis with varimax rotation, KMO = 0.827, was run as before, to clarify the nature of the shared variances between exp FOL, DTV, log block design, marbles, and covariation. A three-factor model provided the clearest solution, with the first factor explaining 33% of variance, the second 28%, and the third 21% (**Table 6**, study 2 rotated component matrix). This confirmed log block design as being most closely related to marbles, and FOL to DTV, but covariation as being distinct.

Taking exponential FOL as standing for both spatial–temporal measures, maximum likelihood path analysis was used to examine whether either the extended mediation model identified in study 1 or a reversed version of this (i.e., with covariation and marbles swapping position) provided an adequate fit to the data. These models were contrasted with a further extension of this, in which log block design and marbles fed directly into FOL alongside covariation, reflecting the somewhat more balanced influence of the two statistical measures. For both causal indices, the further extended model provided the best fit to the data: in each case, $\chi^2 = 0.315$, $p = 0.575$, $df = 1$. **Figure 9** illustrates the model plus path coefficients obtained for overall causal performance.

Again, there was a stable pattern of effects in which non-verbal ability, covariation, and probability all support spatial–temporal analysis, but with each also influencing specific aspects of causal thinking directly. In this sample, FOL and vocabulary have the strongest effects for both causal measures: for inference, the effects of FOL and vocabulary are 0.36 and 0.41, and those of non-verbal ability, probability, and covariation, 0.20, 0.01, and 0.09.

As in study 1, further moderation analyses indicated no interaction effect between log block design, marbles, and FOL in predicting causal performance.

Discussion

Despite differences in the sample, study 2 confirmed the developmental trends observed in study 1 in the ability to analyze probability, covariation, and spatial–temporal information, with if anything clearer increases across the age groups. Once again, the use of all these kinds of information was consistently associated with both overall causal performance and inference, and as before, they appeared to have interrelated influence that was also connected to non-verbal ability.

Marbles was again a significant predictor to begin with in both regression analyses, but its influence dropped with the inclusion of covariation and – in contrast to study 1 – again when block design was added. Covariation was marginally the stronger predictor here and more related to spatial–temporal ability – possibly reflecting the revised measure we used. However, neither statistical measure was a significant predictor in either of the final models, being overtaken in both cases by the spatial–temporal measures, especially FOL. The reduced influence of probability and covariation here is plausibly a reflection of the less developed nature of both competences – and causal inference in this sample. FOL subsumed the influence of DTV, to which it was clearly related, and was the strongest predictor of causal reasoning, alongside verbal and non-verbal ability. The clear impact of verbal ability in this study seems most obviously to be attributable to the more mixed sample, though it cannot in fact be a function of a greater spread of ability, since the variance was actually less than in study 1. Instead, it seems more likely that it reflects vocabulary being a greater influence at lower levels of ability.

In spite of these variations, study 2 largely replicated the results of study 1, while extending them to show that the network of interrelated influences on both spatial–temporal analysis and causal reasoning is based on at least partially unique contributions from probability, covariation, and non-verbal ability. As before, the implication is that statistical and non-verbal ability support spatial–temporal analysis by allowing the capture of patterns of relationship. However, that spatial–temporal ability was a stronger predictor of causal reasoning when statistical and non-verbal ability were less developed and had less direct influence themselves suggests that it is not dependent on these, but rather that each has an independent developmental trajectory – consistent with the factor analysis results.

In line with the results from study 1, the stronger influence of probability on inference and its weak association with covariation suggest that, even in this lower-performing sample, understanding of probability plays a distinctive role in thinking about causal mechanisms, beyond sensitivity to statistical patterns *per se*.

GENERAL DISCUSSION

This study aimed to develop a battery of statistical reasoning tasks, suitable to measure individual differences across a range of developmental levels. It then compared the predictive role of statistical and spatial–temporal analysis in children's causal thinking about continuous natural processes. Across two studies, in total, five statistical tasks were employed. Taking into account the literature, four of them were developed by the authors for this particular project. **Table 7** summarizes the similarities and differences between the two cohorts.

Developmental Trajectories

Children's responses showed clear progress with age on all tasks. The sample characteristics were different in both studies. Study 1 employed higher SES children, while study 2 employed a mixed SES sample and introduced causal tasks with a scientific method approach alongside a modified covariation task. In study 1, there was greater growth on the causal and spatial–temporal tasks between Y1 and Y3 than between Y3 and Y5; in study 2, there were gains on both across the three age groups. On the causal tasks, in both studies, children's inference responses were restricted at all ages, and even in Y5, they seemed to find it difficult to explain the mechanisms mediating cause-effect relationships, focusing on more observable and salient factors and variables. In contrast, performance on the liquid flow task in particular approached ceiling by Y5 in both the five- and six-step versions, regardless of sample differences.

Although children showed some different patterns in different tasks, on the present measures, growth in statistical thinking appeared to be slower than that in spatial–temporal ability, but faster than that in causal inference. Past research examining children's and adult's covariation and probabilistic thinking in causation – Bayesian and causal learning literature – has focused on the identification of the structure of causal relations between distinct variables (e.g., the relationship between the use of aspirin and headache), or the strength of these (e.g., the degree to which aspirin alleviates headaches; see Lagnado et al., 2007) based on summaries of repeated observations, or compared children's understanding of common cause and causal chain structures (McCormack et al., 2015, 2016). Although we do not have data on direct comparability with the more detailed tasks used in the developmental literature, our mini-statistics tasks showed sensitivity to differences in children's statistical thinking that are largely in line with the results from those in the literature.

On the covariation task, children appeared to progress with age in their ability to assess co-occurrences, but to still be refining this further by Y5, especially with respect to numerical quantification. This is in some ways consistent with Shaklee and Mims' (1981) finding, using a causal event-based approach that children's strategies for addressing covariation increased in complexity with age. It appeared that in our tasks, older children performed better on the computation

TABLE 7 | Characteristics of the two cohorts.

	Study 1 (*N* = 107)	Study 2 (*N* = 124)	Significant effects	
			Study 1	Study 2
Hypothesis	Spatial–temporal analysis provides a bridge between observation of continuous processes and their causal analysis			
Sample characteristics	Middle and high SES	Low and middle SES		
Implementation of causal phenomena	A three-stage implementation (prediction, description, explanation)	A five-stage implementation (observation, prediction, justification, testing, explanation)		
Spatial–temporal tasks	Flow of liquid	Flow of liquid	✓	✓
		Distance, time, velocity (DTV)	–	X
Probability and covariation tasks	Randomness	–	X	–
	Marbles	Marbles	X	X
	Cards	–	X	–
	Covariation (computer-based)	Covariation (with cards)	X	X
Verbal task	WASI vocabulary	WASI vocabulary	X	✓
Non-verbal task	WASI block design	WASI block design	✓	✓

of estimates of associational strength. Moreover, although our tasks only focused on co-occurrence (i.e., two cells of the 2 × 2 contingency table), our data also suggest that progress may be slower where the focus is on frequency relationships *per se*, rather than on frequency relations between cause and effect, perhaps because the former are in some sense more abstract.

On both versions of the covariation tasks, one item provided participants with a hundred percent co-occurrence information, which made the interpretation of stimuli unambiguous; the remainder presented incomplete information which increased the ambiguity from 75% in study 2 to 50%, and to 0%. Children mostly seemed to deal well with hundred percent co-occurrence (AB cases, perfect covariation), but less well with the zero correlation, and they had greater difficulty still interpreting degrees of co-occurrence in between with any precision, especially in the youngest age group. This is consistent with Koerber et al.'s (2005) finding that even older children show difficulties in interpreting instances of non-covariation between two distinct events. However, unlike that study, our tasks did not include a conflict between previous beliefs and causal evidence requiring children to test their hypotheses, only to interpret imperfect covariation patterns in a non-causal context. The consistent difficulty in interpreting non-covariation data in both approaches suggests a more fundamental problem that requires further investigation.

In the probability tasks, young children did not show a clear numerical grasp of probability. They did show some understanding of possibilities and their thinking on these tasks seemed to be binary: the majority focused on whether there was a good chance of winning or losing, rather than the degree of that chance. This study therefore agrees with Piaget and Inhelder (1975), White's (2014), and McCormack et al.'s (2015, 2016) findings, adding to those that thinking with numbers and computing probabilities appeared to start from Y3 onward, consistent with the covariation findings, but in neither case did children begin to approach ceiling performance, even by Y5.

There were also departures from past findings. In particular, many of the younger children in the higher performing study 1 sample did not find it easy to make the distinction between predictable and random events, contrary to Kuzmak and Gelman (1986), with our randomness task showing substantial variation around the mean of 1 (with a high standard deviation, it indicates that majority of children were scoring zero) in Y1, and growth coming predominantly between Y3 and Y5. Conversely, they did not find it harder to deal with frequentist probability task which required sampling (assessed by cards task) neither with other probability task (assessed with marbles), as suggested by work on probability learning (e.g., Brainerd, 1981). The cards task exhibited more or less exactly the same developmental profile as the marbles task – though the use of summary presentations may have helped – and the two were strongly correlated. On both tasks, there was a general improvement of probabilistic thinking through the elementary age range from awareness of variation in likelihood to numerically precise calculation of this. While recent cognitive-developmental work has focused on tasks sensitive even to the youngest children's level of skill (Schlottmann and Wilkening, 2011), our tasks stretched even the oldest children in the sample, as intended, especially with respect to proportional calculations – ratios and decimals – as opposed to more basic judgments.

Relationships Between Causal Reasoning, Spatial–Temporal, Covariation, and Probabilistic Thinking

Despite the covariation and probability tasks drawing on related types of frequency information, a distinction between our key predictors was confirmed by both correlational and factor analyses in both studies. In study 1, frequentist cards and other probability tasks (marbles) were closely related as compared with covariation and randomness tasks, and the covariation task appeared to demand a more distinct competence. In study 2, both spatial–temporal measures (liquid flow and DTV) loaded onto the same factor, while marbles and covariation again remained independent, albeit with marbles being associated with block design, once more suggesting that the ability to utilize covariation, probability, and spatial–temporal information requires different competences.

Most people would agree that forms of statistical reasoning, as emphasized by Humean approaches to causality, are useful for causal thinking about discrete events, which lend themselves easily to frequency-based analyses, but the present study showed that this form of thinking also relates to causal thinking about extended processes, which have no perceptually distinct components. One basic possibility is that the role of statistical sensitivity and non-verbal ability is primarily one of enabling forms of pattern detection. Block design assesses the ability to analyze and reconstruct perceptual patterns, which facilitates the detection and representation of causal effects; covariation assesses the ability to track connections, which facilitates the identification of relationships between variables and outcomes; and probability assesses the ability to track the "definiteness" of outcomes, which facilitates awareness of strength of effect, e.g., the relative impact of a variable, in this case on speed of effect. The integration of Kantian mechanism-based and Humean statistical thinking highlighted by our results echoes recent theoretical debates in the literature on causal thinking about discrete events (see Waldmann, 2017). However, our novel contribution is not just the application to continuous processes, but to the correlational, individual difference approach. The same approach might also be useful in the future to study how children's thinking about discrete causal events develops.

While the statistical variables were largely not significant predictors in the regressions, the path analyses, however, found that probability and covariation formed a network of interrelated competences influencing causal reasoning, along with non-verbal and verbal ability. It should be noted that regression models portray the relations from the raw data and cannot provide more sensitive statistics as to, for instance, what is the nature of residuals after each step. We can observe the effects of each variable by assessing the change in beta values after each step. Thus, even if a variable remains non-significant in the final model, the chance in the beta values shows us whether the variable contributed to the model one way or another. In both studies, we captured these widespread interactions with the path analyses. It is the nature of this network that needs to be explained.

Naturally, our interpretations are limited to the task characteristics and statistical methods. We cannot conclude whether or not the basic understanding of, for instance, randomness assists causal inference in continuous processes. Similarly, it is not clear why marbles was a stronger predictor than covariation in both studies. Although we have an idea about the possibilities drawing marbles to lose its predictive power for both causal indices in study 2 (e.g., we had a more powerful covariation task, and there was an additional spatial–temporal variable involved, which reduced the variance explained by both marbles and covariation), these do not analyze the unique nature of the tasks. Moreover, differences between the samples in terms of relative developmental level across the various indices may have played a role. That the spatial–temporal predictors trumped statistical predictors fits with the notion that temporal information overrides covariation when a causal structure needs to be inferred, as in the event literature with adults (see also Waldmann and Holyoak, 1992; Cheng, 1993; Waldmann et al., 1995; Cheng et al., 1996) and with children (see Siegler, 1975; Mendelson and Schulz, 1976; Bullock et al., 1982). Although our focus is on continuity rather than contiguity, the data from both studies show similar outcomes: statistical thinking appears to be promising in terms of supporting reasoning about mechanisms.

Understanding of probability seems to do something more than is captured by this proposed indirect influence on causal thinking. The relationship of performance on the marbles task in both studies to inference of mechanisms and its more distinct predictive power, especially in the higher performing study 1 sample, both suggest that it promotes some other additional insight. A cross-check between probability and causal task performance shows that in both studies children who had perfect marbles scores were more likely to provide high-level inference scores and make reference to mechanisms (n = 21 in study 1; n = 19 in study 2). Probabilistic thinking seems therefore to be important not only for the identification of the strength of the effects of variables, but for considering unseen elements of causal processes. It is plausible that awareness of probability drives a general heightening of sensitivity to the operation of unseen factors, as argued in the introduction. This would be consistent with children exhibiting similar limitations in probability scores and references to mechanism.

However, construction of a dynamic mental representation tying spatial–temporal information together to envisage the operation of *specific* mechanisms still requires in addition a time-based analytical and constructive ability, as captured by the FOL tasks. In other words, non-verbal ability, probability, and covariation help by enabling children to identify variables and to sense that there is more to be explained about how these operate, but as the data suggest, it remains primarily spatial–temporal ability that takes them beyond this to coordination of actual information and ideas of mechanism.

Verbal ability also appears to be necessary to get all of this off the ground, given its influence among the lower-performing sample in study 2. However, all these competences seem to have distinct developmental trajectories and converge on support of causal inference. The nature of the growth of this convergence – and how far it can and possibly needs to be deliberately promoted – requires further investigation.

Two lines of inquiry can investigate this convergence between causal reasoning and distinct competences, one empirical, e.g., experimental studies aiming to elaborate on the aspects of statistical and spatial–temporal thinking, and another methodological, e.g., studies investigating the nature of the causal tasks in relation to intelligence tests. Regarding the first line of inquiry, the present study provided the first dataset, using intelligence measures as controls. It should be noted that these measures have high statistical reliability and explanatory power, and they challenge substantially other tasks present in the same model. One can expect statistical form of thinking to be more predictive in different models when intelligence measures are excluded. In fact, when we did so, the covariation task explained a unique variance in causal thinking (β = 0.178, p = 0.034) along with age and FOL. The increase in beta values of flow liquid was also substantial. This highlights the above interpretation taking into account the nature of shared variances in models, i.e., how the strongest predictor subsumes the beta of others in regression models.

In line with the methodological inquiry, a follow-up study employed three intelligence measures and investigated their relevance to the above causal tasks. The data were analyzed based on the Tucker-Drob (2009) model, which is constructed based on an integration of Horn–Cattell's theory of fluid and crystallized intelligence (Horn and Cattell, 1966). The study found very high correlations between the measures, where general intelligence factor explained about 62% of the variance in causal tasks. This effect was independent of age and the model was able to analyze the nature of the residuals (Dündar-Coecke, under review). This result suggests that there may also be a strong link between spatial–temporal reasoning and intelligence types, which clearly merits further investigations.

CONCLUSION

An important contributor to causal reasoning about continuous processes is spatial–temporal analysis. When its influence is compared with that of statistical reasoning, it remains as the strongest predictor. However, statistical reasoning made both direct contributions and exerted an indirect influence *via* spatial–temporal analysis. The findings here highlight the multiple and complex determinants involved in such thinking. This is the first investigation employing process-based causal tasks to examine the role of covariation and probability alongside spatial–temporal ability. Further studies can explore the unique nature of the tasks and their relations to other forms of reasoning.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

Ethical approval for the studies involved human participants was obtained from the UCL Institute of Education Research Ethics Committee, University College London. Children's verbal responses were also received before the start of each session. Children were not included in testing without their verbal consents even if their parental consent was available.

AUTHOR CONTRIBUTIONS

SD-C conceptualized and developed the idea, conducted the research and analyses, and wrote the original manuscript. AT supervised these processes, verified the analytical processes, and reviewed and edited the manuscript. He was also involved in data and reliability analyses. AS supervised these processes and reviewed and edited the manuscript. All authors discussed the results and contributed to the final manuscript.

FUNDING

This research was funded by an award from the UK Economic and Social Research Council to SD-C (Grant number: ES/P000592/1) and by an internal grant provided by the Institute of Education, University College London.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.525195/full#supplementary-material

REFERENCES

Acredolo, C., O'Connor, J., Banks, L., and Horobin, K. (1989). Children's Ability to Make Probability Estimates: Skills Revealed Through Application of Anderson's Functional Measurement Methodology. *Child Dev.* 60s, 933–945. doi: 10.2307/1131034

Ahn, W., Kalish, C. W., Medin, D. L. and Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition* 54, 299–352. doi: 10.1016/0010-0277(94)00640-7

Anderson, N. H., Schlottmann, A. (1991). "Developmental study of personal probability," in *Contributions to Information Integration Theory: Developmental, Vol. 3,* ed. N. H. Anderson (Lawrence Erlbaum), 111–134.

Bayless, S., and Schlottmann, A. (2010). Skill-related uncertainty and expected value in 5-to 7-year-olds. *Int. J. Method. Exp. Psychol.* 31, 677–687.

Brainerd, C. J. (1981). Working memory and the developmental analysis of probability judgment. *Psychol. Rev.* 88, 463–502. doi: 10.1037/0033-295x.88.6.463

Bryant, P., and Nunes, T. (2012). Children's understanding of probability: a literature review (full report). Nuffield Foundation. Available online at: https://www.nuffieldfoundation.org/wpcontent/uploads/2019/11/Nuffield_CuP_FULL_REPORTv_FINAL.pdf (accessed March 2, 2018).

Bullock, M., Gelman, R., and Baillargeon, R. (1982). "The development of causal reasoning". In *The developmental psychology of time*. (Ed.) W. J. Friedman. New York, NY: Academic Press. 209–254

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychol. Rev.* 104, 367–405.

Cheng, P. W. (1993). "Separating causal laws from casual facts: pressing the limits of statistical relevance". In *The psychology of learning and motivation*. (Ed.) D.L. Medin Vol. 30. New York, NY: Academic Press. 215–264. doi: 10.1016/s0079-7421(08)60298-4

Cheng, P. W., Park, J., Yarlas, A. S., and Holyoak, K. J. (1996). "A causal- power theory of focal sets". In The psychology of learning and motivation 34. (Eds.) D. R. Shanks, K. J. Holyoak, and D. L. Medin. San Diego, CA: Academic Press. 313–357. doi: 10.1016/s0079-7421(08)60564-2

Dennis, M., and Ahn, W. K. (2001). Primacy in causal strength judgments: the effect of initial evidence for generative versus inhibitory relationships. *Memory Cogn.* 29, 152–164. doi: 10.3758/bf03195749

Dündar-Coecke, S., Tolmie, A., and Schlottmann, A. (2019). Children's reasoning about causal processes: the role of verbal and nonverbal ability. *Br. J. Educ. Psychol.* 2019:12287 doi: 10.1111/bjep.12287

Dündar-Coecke, S., Tolmie, A., and Schlottmann, A. (2020). The role of spatial and spatial-temporal analysis in children's causal cognition of continuous processes. *PLoS One* 15:e0235884. doi: 10.1371/journal.pone.0235884

Erdfelder, F. F., Lang, E., and Buchner A. G. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39: 175–191. doi: 10.3758/bf03193146

Ferguson, T. J., Olthof, T., Luiten, A., and Rule, B. G. (1984). Children's use of observed behavioral frequency versus behavioral covariation in ascribing dispositions to others. *Child Dev.* 55, 2094–2105. doi: 10.2307/1129782

Horn, J. and Cattell, R. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *J. Educ. Psychol.* 57, 253–270. doi: 10.1037/h0023816

Huizenga, H. M., Crone, E. A. and Jansen, B. J. (2007). Decision—making in healthy children, adolescents and adults explained by the use of increasingly complex proportional reasoning rules. *Dev. Sci.* 10: 814–825. doi: 10.1111/j.1467-7687.2007.00621.x

Hume, D. (1739/1978). *A treatise of human nature*. Oxford: Oxford University Press.

Griffiths, T., and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychol. Rev.* 116, 661– 716.

Jenkins, H., and Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychol. Monogr.* 7, 1–17. doi: 10.1037/h0093874

Kelley, H. H. (1967). "Attribution theory in social psychology. In Nebraska Symposium on Motivation 15. (Ed.) D. Levine Lincoln: University of Nebraska Press. 192–238.

Kelley, H. H. (1973). The processes of causal attribution. *Am. Psychol.* 28, 107–128. doi: 10.1037/h0034225

Koerber, S., Sodian, B., Thoermer, C., and Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss J. Psychol.* 64 141-152.

Koslowski, B., Okagaki, L., Lorenz, C., and Umbach, D. (1989). When covariation is not enough: the role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Dev.* 60, 1316–1327. doi: 10.2307/1130923

Kuzmak, S., and Gelman, R. (1986). Young children's understanding of random phenomena. *Child Dev.* 57, 559–566.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., and Sloman, S. A. (2007). "Beyond covariation: cues to causal structure". In *Causal learning: psychology, philosophy, and computation*. (Eds.) A. Gopnik and L. Schulz 154–172. England: Oxford University Press. doi: 10.1093/acprof:oso/9780195176803.003.0011

Marsh, J. K., and Ahn, W. K. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *J. Exp. Psychol. Learn. Memory Cogn.* 35, 334–352. doi: 10.1037/a0014929

Mendelson, R., and Schulz, T.R. (1976). Covariation and temporal contiguity as principles of causal inference in young children. *J. Exp. Child Psychol.* 22, 408–412. doi: 10.1016/0022-0965(76)90104-1

McCormack, T., Bramley, N. R., Frosch, C., Patrick, F. and Lagnado, D. A. (2016). Children's Use of Interventions to Learn Causal Structure. *J. Exp. Child Psychol.* 141, 1–22. doi: 10.1016/j.jecp.2015.06.017

McCormack, T., Frosch, C., Patrick, F. and Lagnado, D. A. (2015). Temporal and statistical Information in causal structure learning. *J. Exp. Psychol Learn. Memory Cogn.* 41, 395–416. doi: 10.1037/a0038385

Piaget, J. (1969/2006). *The child's conception of time*. England: Routledge.

Piaget, J., and Inhelder B. (1975). *The origin of the idea of chance in children*. New York,NY: Norton.

Perales, J. C., Shanks, D. R. and Lagnado, D. A. (2010). Causal Representation and Behavior: The Integration of Mechanism and Covariation. *Open Psychol. J.* 3, 174–183. doi: 10.2174/1874350101003010174

Reyna, V.F., and Brainerd, C.J. (1994). "The origins of probability judgement: a review of data and theories". In *Subjective probability*. (Eds.) G. Wright and P. Ayton. Chichester: Wiley.

Schlottmann, A. (2001). Children's probability intuitions: Understanding the expected value of complex gambles. *Child Dev.* 72, 103–122. doi: 10.1111/1467-8624.00268

Schlottmann, A., and Anderson, N. H. (1994). Children's judgments of expected value. *Dev. Psychol.* 30, 56–66.

Schlottmann, A., and Wilkening, F. (2011). "Judgment and decision making in young children," in *Judgment and Decision Making as a Skill*, eds M. Dhami, A. Scholottmann, and M. Waldmann (Cambridge University Press), 55–83.

Schlottmann A., and Wilkening, F. (2012). Judgment and decision making in young children: probability, expected value, belief updating, heuristics and biases. In *Judgment and decision-making as a skill: learning, development, evolution*. (Eds.) M. Dhami, A. Schlottmann, M. Waldmann. Cambridge, UK: Cambridge University Press.55–83.

Schultz, T.R. (1982). Rules of causal attribution. *Monogr. Soc. Res. Child Dev.* 47:194.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., and Jenkins, A. C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. Cognition 109, 211–223. doi: 10.1016/j.cognition.2008.07.017

Shaklee, H., and Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Dev.* 52, 317–325. doi: 10.2307/1129245

Shaklee, H., and Paszek, D. (1985). Covariation judgment: systematic rule use in middle childhood. *Child Dev.* 56, 1229–1240 doi: 10.2307/1130238

Shanks, D.R., and Dickinson, A. (1988). "Associative accounts of causality judgment". In *The psychology of learning and motivation–advances in research and theory*. 21. (Ed.) G.H. Bower, 229–261. San Diego, CA: Academic Press. doi: 10.1016/s0079-7421(08)60030-4

Shultz, T, and Mendelson, R. (1975). The use of covariation as a principle of causal analysis. *Child Dev*. 46, 394–399. doi: 10.2307/1128133

Shultz, T. R. (1982). Rules of causal attribution. *Monogr. Soc. Res. Child Dev*. 47, 1–51. doi: 10.2307/1165893

Siegler, R. S. (1975). Defining the locus of developmental differences in children's causal reasoning. *J. Exp. Child Psychol.* 20, 512–525. doi: 10.1016/0022-0965(75)90123-x

Sobel, D. M., Sommerville, J. A., Travers, L. V., Blumenthal, E. J., and Stoddard, E. (2009). The role of probability and intentionality in preschoolers' causal generalizations. *J. Cogn. Dev.* 10, 262–284. doi: 10.1080/15248370903389416

Tucker-Drob, E.M. (2009). Differentiation of cognitive abilities across the lifespan. *Dev. Psychol.* 45, 1097–1118. doi: 10.1037/a0015864

Waldmann, M. (2017). The Oxford handbook of causal reasoning. New York, NY: Oxford University Press.

Waldmann, M. R., and Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *J. Exp. Psychol. Gen.* 121, 222–236. doi: 10.1037/0096-3445.121.2.222

Waldmann, M.R., Holyoak, K.J., and Fratianne, A. (1995). Causal models and the acquisition of category structure. *J. Exp. Psychol. Gen.* 124, 181–206. doi: 10.1037/0096-3445.124.2.181

Wechsler, D. (2011). Wechsler abbreviated scale of intelligence second edition (WASI-II). San Antonio, TX: NCS Pearson Education.

White, P. (2014). Singular clues to causality and their use in human causal judgment. *Cogn. Sci.* 38, 38–75. doi: 10.1111/cogs.12075

Wilkening, F. (1981). Integrating velocity, time and distance information: A developmental study. *Cogn. Psychol.* 13, 231–247. doi: 10.1016/0010-0285(81)90009-8

Xu, F., and Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition* 112, 97–104. doi: 10.1016/j.cognition.2009.04.006

Conflict of Interest: 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership