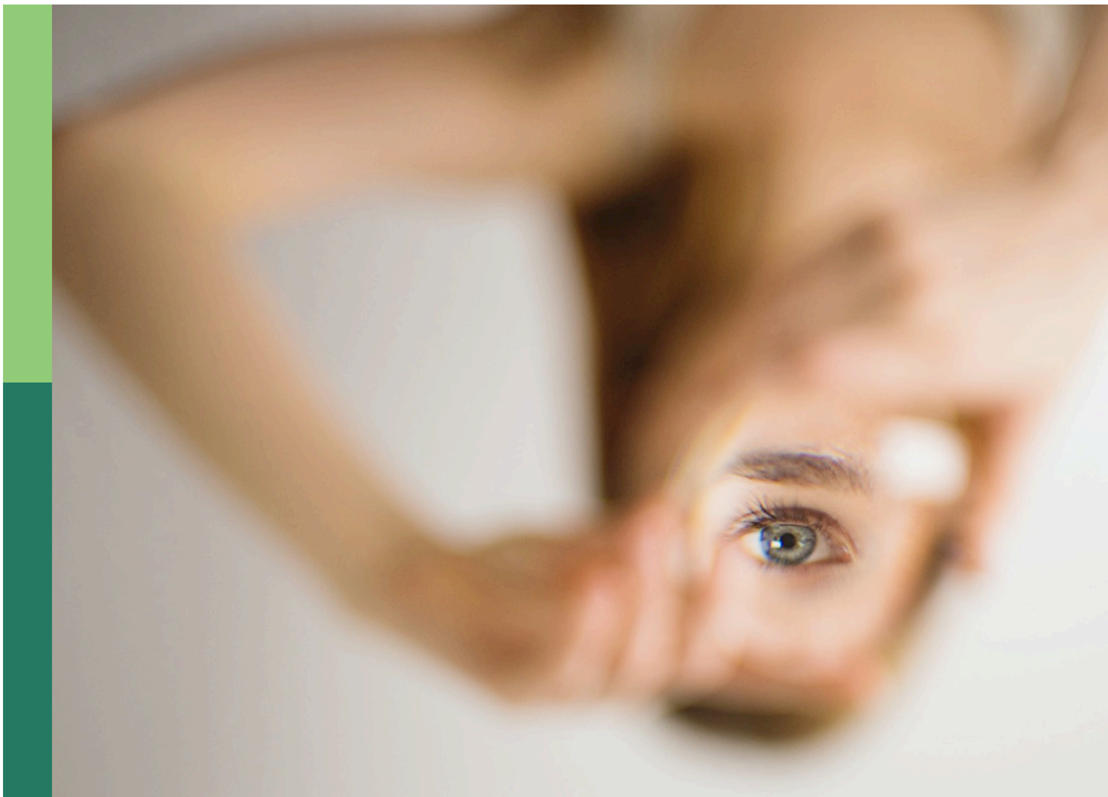
frontiers | Research Topics

Insights in: Theoretical and philosophical psychology

Edited by
Anna M. Borghi, Luca Tummolini, Guy Dove
and Chiara Fini

Published in
Frontiers in Psychology



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-3675-9
DOI 10.3389/978-2-8325-3675-9

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

Insights in: Theoretical and philosophical psychology

Topic editors

Anna M. Borghi — Sapienza University of Rome, Italy
Luca Tummolini — National Research Council (CNR), Italy
Guy Dove — University of Louisville, United States
Chiara Fini — Sapienza University of Rome, Italy

Citation

Borghi, A. M., Tummolini, L., Dove, G., Fini, C., eds. (2023). *Insights in: Theoretical and philosophical psychology*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-3675-9

Table of contents

Frontiers in Psychology

TYPE Editorial
PUBLISHED 26 September 2023
DOI 10.3389/fpsyg.2023.1268864

OPEN ACCESS

EDITED AND REVIEWED BY
Edoardo Datteri,
University of Milano-Bicocca, Italy

*CORRESPONDENCE
Chiara Fini
✉ chiara.fini@uniroma1.it

RECEIVED 28 July 2023
ACCEPTED 01 September 2023
PUBLISHED 26 September 2023

CITATION
Fini C, Tummolini L, Dove G and Borghi AM (2023) Editorial: Insights in: theoretical and philosophical psychology.
Front. Psychol. 14:1268864.
doi: 10.3389/fpsyg.2023.1268864

COPYRIGHT
© 2023 Fini, Tummolini, Dove and Borghi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Insights in: theoretical and philosophical psychology

Chiara Fini[1]*, Luca Tummolini[2], Guy Dove[3] and Anna M. Borghi[1,2]

[1]Department of Dynamic and Clinical Psychology, and Health Studies, Sapienza University of Rome, Rome, Italy, [2]Institute of Cognitive Sciences and Technologies, Italian National Research Council, Rome, Italy, [3]Department of Philosophy, University of Louisville, Louisville, KY, United States

KEYWORDS

embodied cognition, methodological and theoretical approaches, decision making, health, computational modeling

Editorial on the Research Topic
Insights in: theoretical and philosophical psychology

Multifaceted reflections across different domains of knowledge—ranging from the philosophy of science and new interdisciplinary theoretical backgrounds in the field of psychology and computational neuroscience, to social perception, health and decision-making processes, are objects of the current Theme Issue.

The Theme Issue revolves around three nodes which are brought to the attention of scientists and philosophers as timely issues to deal with:

i) the necessity of rethinking theoretical and methodological practices in social and life sciences in conformity with the natural evolution of the domains of knowledge,
ii) the enlargement of (social) embodied perception field by including also psychopathological conditions,
iii) the definition of the theoretical and ethical borders about the concept of human health and the associated decision-making processes.

The first thematic node is related to the relevance acquired within the scientific community of Open Science practices (e.g., preregistration). In this regard, Jacobucci questions the use of confirmatory and exploratory labels in the era of big data. The author argues that, after the replication crisis, in psychology, confirmatory research is becoming more frequent and increasingly requested. At the same time, the advent of big data leads to the frequent incorporation of exploratory elements. The author argues that applying the simple labels "confirmatory" and "exploratory" can present several limitations and that only the simplest studies can be considered as really confirmatory. Describing their research as confirmatory, researchers tend to hide uncertainty in their theoretical foundations. Overall, according to the author, using the label confirmatory and exploratory has many drawbacks. Instead, the author argues for avoiding the use of the rigid labels of confirmatory and exploratory because they are out of date in a time in which Hypothetic-Deductive research is becoming less frequent. Instead, he advocates for the necessity to explain in a detailed but more flexible way "how replication/generalizability was addressed statistically, the form of reasoning used in developing the study procedures, whether explanation, prediction, or description is the primary aim, and finally, what stage of theory generation, development or appraisal the research line is in."

Flanked by such aspects, another critical issue is the necessity to build solid theoretical background to be dis(confirmed), allowing consistent advances in knowledge. The reproducibility crisis which has plagued the behavioral sciences in the last years has prompted the development of new scientific practices to avoid repeating the same mistakes of the past. However, in addition to the adoption of the many transparency measures which are intended to improve the quality of data, Witte et al. argue for the importance of developing complementary methods to improve our theory construction. To this end, they propose and assess a new method to evaluate the similarity between a theoretically predicted effect and observations which improve the identification of the underlying theoretical construct. Scientific progress needs to rely on these complementary approaches.

A different approach is explored in Teo's article, which adopts the lens of "white epistemology," a core tenet of Critical Race Theory—CRT—to argue that psychological science turns out to draw on a race-biased research practice. The article is articulated around a main argument: the impossibility of relying only on the goodness of the scientific method to declare that a research practice has brought clear, reliable, and interpretable outcomes. When treating humanities and psychological issues, epistemological contexts and temporality—which offer the necessary background to interpret psychological differences among social groups—can't be ignored. Results obtained by applying a rigorous scientific method are meaningful only within a sociocultural scaffolding; otherwise, the same results might be erroneously interpreted as underlying "objective" or even worse "eugenetical" differences among populations. As has happened in the past, research outcomes in sociocultural domains—not correctly contextually framed—have offered a scientific justification for social stigma toward some social groups. In conclusion, the de-contextualized interpretation of outcomes obtained with the scientific method, without an appropriate epistemic complexity is not appropriate when studying humans and races.

Regarding the scientific content, instead of the practice, contemporary psychological/neuroscientific knowledge is evolving toward increased cross-field contaminations, with a rapid growth of intertwined hybrid disciplines.

Inspired by the confluence of many diverse approaches into the coherent subdiscipline of robophilosophy, Krageloh et al. make a persuasive case for pushing a similar development in psychological science as well. Given the breakthroughs in AI and robotics and their impact in so many diverse domains, they propose that it is time to give rise to the new field of "robopsychology." A robopsychology may help organize ongoing streams of research that explore both the impact of these technological innovations on human minds as well the way in which these artifacts may acquire a mind of their own. The rise of a "psychology *of*, *for*, and *by* robots, robotics, and artificial intelligence" is surely a topic that needs to be widely discussed by our community. Regarding the evolution of theoretical approaches in neuroscience, de Wit and Matheson posit as a sensitive topic the re-conceptualization of the best modality of functional mapping. While a weak contextualism, allows to stay at a very abstract level of explanation about a brain area's function, a strong contextualism is open to re-classifications and embraces the context-dependent frame to understand, test and map all the neuro-cognitive mechanisms. Context-dependent neural tuning, neural reuse, degeneracy, plasticity, functional recovery, and the neural correlates of enculturated skills each show that there is a lack of stable mappings between organismal, computational, and neural levels of analysis. Following the authors' perspective, each attempt of mapping discrete neuro-cognitive mechanisms, at neural, computational and phenomenological level, is not feasible. Indeed, recent research shows that behavioral goals and contextual variables affect neural recruitment. A re-conceptualization in cognitive neuroscience about the best modality of functional mapping, appears as necessary. Finally, Ahmad et al. identify promising strands of the Social Exchange Theory—SET—, an inspiring approach to frame social behavior for multidisciplinary domains like social psychology, sociology, anthropology and management science. They assessed the state of the art in the field and developed a systematic approach to identify the most promising directions for future research in SET, which, according to the authors, should move beyond the role of positive reciprocity exchanges. Hopefully, also thanks to their proposal, this long enduring framework will be able to inspire further studies also in the next future.

The second thematic node refers to the theoretical evolution of the (social) embodied perception field. Kim and Effken present a conceptual analysis in which they connect the disturbance of the ecological self and impairments in the perception of affordances. They illustrate the notion of affordance as introduced by Gibson, and argue that when in the presence of affordances, accomplishing successfully intended actions is a sign of autonomy and control in individuals. Without the capability to perceive and actively respond to affordances, the environment stops being meaningful. The authors propose an indirect way to test and validate the notion of affordances, i.e. referring to individuals with mental disorders, and specifically with disorders derived from disturbance of the minimal self (e.g., schizophrenia, post-traumatic stress disorder, and Alzheimer's disease). They characterize this minimal self as "ecological self," the first form of self we experience in infancy. Following Gibson, they propose that if the perception of self is disturbed, then the ability to attune to exteroceptive information in the environment will be disturbed too. They conclude that impairment in affordance perception might be associated with a disturbance of the self.

Within the psychopathology cluster, eating disorders are particularly prone to be investigated through the lens of (social) embodied perception, as Tramacere suggests in her proposal. Starting from the assumption that (i) when looking at the face of another, the same mirroring circuits–MNS– involved when looking at our own face, are activated, and that (ii) our perception of their face is affected by our feelings toward them, the author contends that it is likely that feelings toward ourselves affect our responses to the mirror image. Thus, our body image would be shaped and represented as a function of our own feelings toward ourselves. In relation with the spontaneous sensorimotor resonance triggered by the other's observation, taking up from the Stern's (2010) notion of the vitality of forms—which capture the expressive style of our actions—Liu et al. propose that this theoretical notion helps explain how we are able to perceive the intentions behind the actions of others. More broadly, the vitality of forms serves as a background condition for our understanding of their mental states.

Then, the third node focuses on health and decision-making processes. Firstly, Binder seeks to reconceptualize how existential suffering is viewed in Western culture. He proposes that it would be better to adopt a concept of "existential health" and thus, to abandon the medical model of pathological suffering. Directly related to the theme of health, Berens and Kim deal with the topic of the controversial debate about the nature of decision-making processes in clinical practice. Specifically, the authors, through a review, list arguments supporting or not the theoretical perspective of risk-assessment decision-making—the idea that the higher the risk involved in a decision, the greater the decisional abilities required for DMC—RS-DMC. In conclusion, most positive defenses of RS-DMC rely on its intuitive appeal, while most criticisms are driven by concern about paternalism or the asymmetry between consent and refusal. Much research about the topic is needed. Finally, a new mathematical model in the Markov process is proposed to explain decision-making dynamics by Bizzarri et al.. The novelty of the proposal relies on the integration of concepts like: tacit knowledge—Pascal's "esprit de finesse"—intuition, emotions, awareness and self- awareness to explain the decision-making processes. Crucially, the obsolete dichotomy between analytical and intuitive (holistic) reasoning is definitely overcome in these mathematical formulations, where both emotional and more implicit factors contribute to decision-making processes. Through the model simulations, it is found that awareness emerges as a dynamic process allowing the decision-maker to switch from habitual to optimal behavior, resulting from a feedback mechanism of self-observation. Furthermore, emotions are embedded in the model as inner factors, possibly fostering the onset of awareness. Importantly, the impact of emotions is re-thought with an explicit dependence on the level of awareness of the individual, so that, the conception that emotion is a noise to be filtered is mitigated by the consideration that it is true at a low state of awareness, and can thus be enhancing for aware individuals. In keeping with this, from a completely different perspective, Kam declines in psychoanalytic terms the following principle: through the "ego inflation" people can take the decision to rationally avoid potential detrimental knowledge and thus to preserve mental wellbeing.

In conclusion, the Theme Issue develops across different dimensions, with the goal to inspire thoughts, ideas, and reflections about methodological and theoretical renewal and progress in the research.

Author contributions

CF: Writing—original draft. LT: Writing—original draft. GD: Writing—original draft. AB: Writing—original draft.

Funding

This project was supported with funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE8 - Project Age-It: Ageing Well in an Ageing Society. This resource was co-financed by the Next Generation EU [DM 1557 11.10.2022].

Acknowledgments

We would like to thank all the authors who participated in the Research Topic.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Stern, D. N. (2010). *Forms of Vitality: Exploring Dynamic Experience in Psychology, the Arts, Psychotherapy, and Development.* New York, NY: Oxford University Press.

PERSPECTIVE
published: 27 January 2022
doi: 10.3389/fpsyg.2022.803792

Suffering a Healthy Life—On the Existential Dimension of Health

*Per-Einar Binder**

Department of Clinical Psychology, Faculty of Psychology, University of Bergen, Bergen, Norway

This paper examines the existential context of physical and mental health. Hans Georg Gadamer and The World Health Organization's conceptualizations are discussed, and current medicalized and idealized views on health are critically examined. The existential dimension of health is explored in the light of theories of selfhood consisting of different parts, Irvin Yalom's approach to "ultimate concerns" and Martin Heidegger's conceptualization of "existentials." We often become aware of health as an existential concern during times of illness, and health and illness can co-exist. The paper discusses how existential suffering in Western culture is described, to an increasing degree, as disorders or psychological deficits, and perfectionistic health goals easily can become a problem. We seek to avoid suffering rather than relate to it, with all the tension that may create. The paper argues that suffering is an unavoidable aspect of people's experience of their lives, and actively relating to suffering must be regarded as a fundamental aspect of health. The need and usefulness of a concept of "existential health" is discussed.

Keywords: existential health, existential concerns, definitions of health, medicalization of life, suffering

OPEN ACCESS

Edited by:
Guy Dove,
University of Louisville, United States

Reviewed by:
Piers Worth,
Buckinghamshire New University, United Kingdom
Frank Martela,
Aalto University, Finland

****Correspondence:***
Per-Einar Binder
per.binder@uib.no

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *28 October 2021*
Accepted: *05 January 2022*
Published: *27 January 2022*

Citation:
Binder P-E (2022) Suffering a Healthy Life—On the Existential Dimension of Health. Front. Psychol. 13:803792. doi: 10.3389/fpsyg.2022.803792

INTRODUCTION

Health is usually what we expect, and it is certainly not a problem. Therefore, health is not necessarily something we reflect about when it is present, and it is often a silent and unnoticed phenomenon (Gadamer, 2018). When we try to conceptualize what health is, a negative definition—absence of illness—is often what comes most immediately to mind. Our need for reflection and story-making arises when we are in trouble, when we have a problem to solve, or something unexpected occurs, such as when illness shows up (Kirmayer, 2000).

Mostly, we become aware of the existential dimension of health during times of illness. When we are ill, we more easily become aware of the finite nature of our being-in-the-world (Yalom, 1980). Illnesses might bring limitations to the activities in life that engage us and give direction. Often illnesses heighten our awareness of our mortality (Kissane, 2012). The uncertainty connected with illness also may also confront us with the undetermined nature of our existence. It challenges us with the fact that the choices we have made so far do not constitute an essence of who we are. Often, our sense of freedom and responsibility feels troublesome to us. This can be mixed with sorrow when we use our freedom to let go of some aspect of our lives. However, sometimes it also opens up the freedom to choose a topic and style for a new chapter in one's life story, not seldom with an unexpected twist (McAdams and Bowman, 2001). A disease can make us reconsider our projects and roles, and demand that we make new choices and priorities. And this process, can be healthy. Illness and a healthy, heightened, and existential awareness can co-exist.

The aim of this paper is to examine the existential context of physical and mental health. The definitions of health made by Hans-Georg Gadamer (2018) and the World Health Organization will be used as a starting point. They will be discussed in light of perspectives on suffering (Schneider, 1999; Miller, 2005), and selfhood consisting of different parts (Bromberg, 1996). Then I will examine the relationship between health and existential concerns through the approaches of Irvin Yalom (1980) and Martin Heidegger (1957), and Ola Sigurdson's (2016, 2019) definition of "existential health." Medicalized and perfectionistic ideals for health are critically discussed, considering the existential conceptualizations. I will propose that being aware of suffering and actively relating to it is part of living a healthy life, and I will also discuss whether a concept of existential health is needed.

WHAT IS HEALTH?

When we are healthy, other aspects of life than health in itself seems more important. We are busy living. Gadamer (2018) describes this as "the enigma of health": our health is of uttermost importance to us, but at the same time, it is not something we usually reflect upon or examine through introspection. When we are healthy, we are in a "condition of being involved, of being in the world, of being together with one's fellow human beings, of active and rewarding engagement in one's everyday tasks" (p. 13).

When we are engaged in our everyday world in a relatively undisturbed way, we feel that this world has what Heidegger (1957) describes as a "homely" quality. Fredrik Svenaeus (2019) explains that illness might transform our experience of everyday life and get with it an "unhomelike" being-in-the-world. Illness might feel like something unfamiliar or alien intruding, either in the body or the mind. Illnesses quickly also penetrate much of our emotional lives. It requires us to change how we imagine the near and sometimes far future. Also, not bringing attention to illness takes effort. As Svenaeus points out, diseases may make us even more preoccupied with going on with our everyday life matters, simply because these matters become hard and painful to accomplish. We find parallels to this line of thinking in Joseph Sandler's (Sandler, 1960) concept of the "background of safety" as a contrast to the world of the traumatized person, where the world becomes a "background of the uncanny." Health is part of our sense of homeliness and safety in the world.

Gadamer (2018) also describes "well-being" as an aspect of health. According to Gadamer, this type of well-being, of greatest value to our lives, mostly escapes our awareness. This well-being is not something that draws our attention in itself, because it is a condition "of being unhindered, of being ready for and open to everything?" (p. 73). But might health as well be more than a pre-reflective phenomenon, working in the background when we are busy living our lives?

The World Health Organization goes some steps further to get a positive conceptualization of health. It defines it as a "a state of complete physical, mental and social well-being and not merely the absence of disease and infirmity." The use of the word "complete" may be interpreted in different ways. On the one hand, the WHO infer a balance of attention paid to overall health's physical, mental, and social parts. On the other hand, the word "complete" can allure to see health as an ideal state, out of reach for most of us, perhaps except in the luckiest and happiest moments of our lives.

Paradoxically, one can argued that "complete" well-being in an idealized sense of the word would make us sick in the long run. Our immune system as well as our musculoskeletal system and our emotional and cognitive abilities are strengthened and keep functioning through stresses and challenges. You need to catch a cold sometimes to keep your immune system strong. People who experience meaning in their life are also often the people who experience a certain amount of stress and worry (Baumeister et al., 2013). Health is not a thing, or something that we possess, but a way of living. This is also something the WHO has taken into account, and in 1986, the WHO also expanded health to encompass "resources for everyday life." In this way, the organization emphasized resilience and the active part of health through dealing with life's challenges.

When Gadamer (2018) describes health as a condition of both being unhindered, and ready for and open to everything, this is also certainly much of an ideal state. Life is full of hindrances, there is often a lot going on that we are not especially ready for, and certainly we often close down rather than open up. We experience huge contrasts in life. Therefore, it can be argued, the experience of being unhindered, ready, and open is something we often notice, even reflect on. These are moments when we feel vitality, and often joy. We might feel grateful for our joys because we are aware of the place suffering has in our life as a whole (Vaillant, 2008). Perhaps health has more to do with the ability to be present in the contrasts of life? Furthermore, perhaps health also has to do with our ability to handle suffering, as an unavoidable fact of life?

The title of Jon Kabat-Zinn and Hanh's (2009) book *Full catastrophe living* signals that all lives also have a "catastrophic side," and that it is possible to stay present in this mix of catastrophe and suffering, and joy and growth that a human life offers. Anger, anxiety, shame, guilt, envy, and sorrow are messengers of bad news. These emotional states are not a trick evolution has made to bother us; they are messengers about realities in our life that we need to address. We suffer when we experience something hindering us, when we feel lonely, or that we lack something, when we feel rejected or threatened, and when we have lost something of great importance to us. Negative emotions have an important place in a rich and meaningful life (Parrott, 2014).

When something is unpleasant and painful, in current culture we often go to the language of illness, and seek explanations in terms of dysregulation or disorders. As Kirk Schneider (1999) points out, when "dysregulation of the emotional brain" (p. 109) is to be described quite unambiguously in mechanical and quantitative terms, suffering is a multifaceted problem relating to dilemmas in existence itself. Suffering entails an alteration, and sometimes a rupture in our life – it is a messenger of change. Illnesses sometimes cause such a rupture, and suffering can be strongly related to a disorder. Important knowledge of high relevance to treatment might be systemized through diagnostic terms. Although this knowledge can be helpful, it also brings a danger of changing our language for the painful challenges in life. When disorders, understood in diagnostic terms, are something to eliminate, suffering is something to deal with. This is also a central part of healing processes. As Ronald B. Miller (2005) points out, psychotherapy primarily deals with suffering. Surprisingly, this is often an unaddressed fact. In an increasing degree, clinical psychology and psychotherapy has been put under pressure by insurance companies and national health services to adapt to the diagnostic language of disorders when making a treatment plan. When it comes to the conceptualization of the patient's problem, clinical psychology and psychotherapy lose something crucial if the medical vocabulary of "disorders" comes in the way of relating to human suffering in an existential sense. Diagnosed "disorders" are abstractions at a considerable distance from the personal life of meaning and importance. Suffering is something humans have in common, and therefore it can be shared. Patients' experience of therapists' empathy is ranked among the most important factors of change in psychotherapy (Norcross, 2011). When psychotherapy works, shared experience of suffering is a fact. The therapist can sense some of the patient's pain, both in their imagination and their body. Moreover, the patient becomes able to relate to their own suffering through the recognition and compassion offered by another human being. There are paths to health that do not have to do with "elimination of illness," but by relating to suffering in friendly, caring, and accepting ways, both in others and in oneself (Esdaile et al., 2021).

Based on a theory of selfhood consisting of multiple parts, Phillip Bromberg (1996) states that "health is the ability to stand in the spaces between realities without losing any of them—the capacity to feel like one self while being many" (p. 166). This need to relate to different parts of self, while being a whole person at the same time, is also in line with what Donald Winnicott (2014) described when he stated that "we are poor indeed if we are only sane" (p. 150).

George Vaillant (2008) describes how joy allows us acknowledge and relate to suffering. A part of my self might feel sorrow over a loss, another part is engaged in a very meaningful project. Both parts are authentic and true, and the sum is bittersweet. I might suffer from a chronic and sometimes painful disease, and at the same time fall in love, and feel that life is opening up deeply; one part of me is in pain, another is on the wings of love. If I can stay present with such contrasts, and they are allowed to be part of me and my life—if I can handle them and feel agency—perhaps this is the highest possible level of health that I can reach? To illuminate this question, we need to examine health as an existential phenomenon.

HEALTH AND EXISTENTIAL CONCERNS

Irvin Yalom (1980) describes four major "ultimate concerns": death, meaninglessness, isolation, and freedom. He describes these as "givens of existence," or an "inescapable part" of being human, and that every person must come to terms with these concerns through active choices to realize their individual potential. These ultimate concerns have common roots with what Heidegger (1957) describes as "existentials," which are structures that form human experience, and that are rooted in our ontological being. The fundamental basis in this structure is "Caring" (*Sorge*), a quality of engagement in the world. "Understanding" (*Verstehen*), "Being-with" (*Mit-Sein*), "Being-toward-death" (*Sein-zum-Tode*), and "Mood" (*Befindtlichkeit*) are other examples of existentials. In our discussion of the existential dimension of health, Yalom's description of the psychological and phenomenological aspects of "ultimate concerns" might be more directly relevant than the more wide-ranging concepts on an ontological level.

However, a problem with Yalom's (1980) conceptualization of ultimate concerns is that they highlight only one of two polarities. He intends to explore a polarity by addressing the pole we tend to resist. For example, when he describes meaninglessness as an existential concern, he uses this as a point of entry to explore the role of meaning-making. Creating and discovering meaning in one's own personal way is necessary to live in an authentic and fulfilling way. As Heidegger (1957) points out, in our everyday mode of experiencing the world, we experience the world as a place where meaning already exists. We are involved in our everyday goal-related tasks in a way that makes us understand them (cf. Heidegger's term "Understanding"). However, existential anxiety is a mood that disrupts our involvement with the familiar signifiers of the world, wipes away the intelligibility we take for given. In this way it also confronts us with meaninglessness. In this way, existential anxiety calls for us to reorient ourselves and commit ourselves to engagements that create new meaning. A possible modification of Yalom's concepts that might be useful in our discussion is to describe these existential concerns as polarities:

(1) Death—and awareness of living a life of one's own.
(2) Meaning—and meaninglessness.
(3) Being-with—and isolation.
(4) Freedom—and limitations and conditionings.

Discussing the existential dimension of health, "embodiment and emotional being" is relevant as a fifth concern; health becomes an issue for us since we are bodily beings. Neither Yalom nor Heidegger explicitly discusses this, however, Heidegger signals that an emotional aspect, "Mood" *(Befindtlichkeit),* is also a way of being that gives us access to our world. Maurice Merleau-Ponty (1962) more actively addresses the fact that our-being-in-the-world is embodied, and Eugene Gendlin (1982) discusses how this embodied mode of being also produces a felt sense of emotional meaning in us. We can be immersed in our bodily felt experience and witness and reflect on these experiences through our capacity for awareness. Our embodied and emotional being has both a proactive and receptive side. Strength and agency on the one side, and vulnerability and receptivity on the other, are polarities connected to embodiment as an existential concern.

Both mental and physical health involves our existential concerns. Our worries about health provide a clue to what types of concerns are involved. Worries about physical illness are often triggered by death-anxiety and the fact that we are embodied beings (Solomon et al., 2015). The strange-looking mole, the uneven heartbeat, the inexplicable tiredness—they easily trigger inner scenarios with death as a possible outcome, and sometimes not without reason. Worries about physical illness, as well as mental illness, can also be connected to concerns about loss of freedom, isolation, and meaninglessness. Worries can both be an invitation to constructive action, and a source of avoidance and passivity (Sweeny and Dooley, 2017). After we have taken the relevant action, death, the possibility of unfreedom, isolation, and meaninglessness are still a part of our lives, and we are still living in a vulnerable and sometimes emotionally sensitive body. In even the healthiest of lives, suffering is lurking around the corner.

If we are to paraphrase Bromberg (1996), we might say that health depends on the ability to "stand in the spaces" (p. 166) between existential polarities:

(1) We can relate to the fact that we will die, and use this insight in a way that enriches our being (Heidegger, 1957); for instance, insight into this fragility of being can invite us to make life choices in line with what we experience to be most important in life (Jaspers, 2010).

(2) We cannot extract meaning from the universe in itself. Meaning is something we create, collectively through language and culture, and individually, through our everyday engagements with other human beings and goal-related tasks with objects in our world (Heidegger, 1957). Through works of art, creativity, and play, we transform our experience of the world and find new meaning (Winnicott, 1971; Heidegger, 2000). Through engagements in activities and projects that we value and that point to something larger than ourselves, we create meaning in our lives (Frankl, 2004). Since meaning depends on our own efforts, meaninglessness and lack of purpose are always a threat becoming apparent through ruptures in our everyday engagements and existential anxiety (Tillich, 2000).

(3) We can relate to the fact that we live lives deeply connected to the destiny of other people and that life is fundamentally relational (Heidegger, 1957; Buber, 2003). At the same time parts of ourselves and our lived experience is out of reach from others, and therefore isolated (Winnicott, 1965). We also recognize that there are parts of other persons' lived experience that we never can directly know. Both being-with and isolation are parts of our lives.

(4) We have a fundamental sense of freedom and responsibility (May, 1999). At the same time, we can relate to the fact that we are limited by habitual ways of being that we are often only dimly aware or unaware of, and that social structures and power limit what choices we see as possible. In all lives, people sometimes must struggle to make the choices they want and need to take.

(5) We can experience ourselves in a body that is felt to be our own, and sensing this body make us feel grounded (Winnicott, 1954). All types of human experience involve a touch-like sense of belonging, also when the content of our feelings are not about the body, but about events in the world that we are part of Ratcliffe (2009). Through the body, we can relate to sensations and emotional states that tell us about what is going on in our world, both at the physical and relational level (Gendlin, 1992). Both agency and receptivity are essential aspects of our bodily involvement (Merleau-Ponty, 1962). This embodied being-in-the-world implies strength and vulnerability as polarities in ourselves, sometimes playing together, sometimes in conflict with each other.

When it comes to health, being-with might be regarded as the most important point of entry, also to be able to reach the other existential dimensions of health. As Daniel Siegel (2010) points out, "feeling felt" is a precondition for psychological growth, especially critical for attachment in infancy, but also for developmental processes later in life. "Feeling felt" describes the ability of one person to encounter another person empathically and authentically on an emotional level. When feeling felt, we intuitively grasp that this other emotionally recognizes our emotional experiences. We feel the emotional resonance. When the others communicate how they perceive us as experiencing persons, this builds and strengthens our sense of existing and being alive. Feeling felt creates a fundamental sense of safety. This emotional bond is a primary mode of being-with, and of utmost importance for being able to handle suffering in life, when we need to create meaning, make choices, when we struggle to be grounded in our bodies, and when we face death.

CONCEPTUALIZATIONS OF EXISTENTIAL HEALTH

Does this way of understanding the existential dimensions of health imply that we need a distinct concept of "existential health"? Several authors have proposed this.

One way of conceptualizing existential health places it as a specific dimension of health in addition to other dimensions. Valerie DeMarinis (2008) describes this existential dimension as a focus on the individual's understanding of existentiality and the way he or she creates meaning. Her concept of existential health is closely linked to participation in meaning-making systems, including those of spirituality and religion, and the ways of relating to rituals and symbolic expressions that are part of such systems. Melder (2012) develops a similar conceptualization, describing existential health as a distinct sphere in addition to the physical, mental, social and ecological, and elaborates on the psychological and creative dimension. She uses Winnicott's theory of the root of creativity in childhood playfulness to explore how the individual projects existential needs into the world, and makes interpretations of it, introjecting from experiences of factors in the outer world that have existential significance, such as religions and philosophies. In her view, the existential health sphere is the sum of such existential meaning-making processes.

Sigurdson (2016) proposes another way to understand existential health. He suggests "existential health" to be the reflexive experience of health. This experience has an intentional quality in how the person relates to their ailment and health. The experience also must have a personal quality; the person relates to this experience as theirs. In this way, he distinguishes between existential health and concepts that relate to more specific dimensions of health, such as spiritual health, and physical, mental, and social well-being. Although his conceptualization mainly addresses the reflexive dimension, Sigurdson (2019) also discusses how embodiment and the way we relate to suffering is a central aspect of existential health. He proposes that we achieve health through learning how to suffer. Although suffering can be seen as a result of our vulnerability as bodily and emotional beings, Sigurdson does not regard it as a passive sensation. Suffering is an active work of acknowledging pain or misfortune, "establishing a kind of solidarity, and learning how to endure it" (p. 96). Suffering means to take an intentional and proactive stance toward painful psychological or bodily realities rather than passively reacting to them. In this sense, the act of suffering and existential health is essentially one and the same.

When DeMarinis and Melder puts existential health closer to the spiritual dimension of health, Sigurdson describes it as a more overarching concept, relating to all aspects of health. The way he connects existential health to embodiment and the ability to "suffer" make his conceptualization more in line with the discussion in this paper. I would add that we need a concept of being simply "aware" in addition to being "reflexive" when it comes to the existential dimension of health, as our relationship to health does not have to be discursive. Our state of being-in-the-world can be also recognized in a more "silent," open, receptive, and at the same time friendly way, as in mindfulness-meditation (Kornfield, 2017). However, in the context of this discussion, this is not the most important distinction.

Do we need existential health in addition to other conceptualizations, or are we more in need of an existential reexamination of the meaning of "health" and its place in life? This does not have to be an either/or, but in a society where health has turned out to be a life goal on its own, this is an important issue to discuss.

THE MEDICALIZATION OF EXISTENTIAL REALITIES

Medical terms are creeping more and more into our everyday language. Experiences of sorrow easily becomes "depression," overwhelming experiences becomes "trauma," fear of failure becomes "anxiety." Concepts from psychology have also crept into our everyday language and the way we handle our life (Madsen, 2014). For instance, we are a culture that seem obsessed with the idea of "self-esteem," tending to see the experience of self-worth as decontextualized from our real efforts and social engagements (Baumeister et al., 2003). Self-esteem has become something we think we possess inside ourselves, which we need to build, and fear losing.

On the one side, our existential suffering, to an increasing degree, is in danger of being described as disorders or psychological deficits (Albarracin et al., 2015). On the other side, one can also argue that there is a cultural trend where we can easily be invited to pursue quite perfectionistic health goals (Fugelli, 2006; Sirois, 2016). In medicine, such health goals may lead more people to be "patients," whereby they receive treatment for potential risks or plastic surgery for normal variants of bodily appearance (Suissa, 2008; Brownlee et al., 2017).

As our lives become medicalized in this way, there is also a danger that we become foreign to them (Miller, 2015). Our troublesome mental states may become less familiar to us, and become the domain of specialists in clinical psychology and medicine. Instead of reflecting our life, our personal struggles, and our values, mental states then gradually came to belong to the categories of health and unhealth. As Patrick Whitehead (2019) points out in his book *Existential health psychology*, treatment always transforms us as people—a bit of us becomes medicalized.

In the strict medical sense, good health is in itself highly compatible with existential health. As the second half of life gives rise to several challenges to both physical and mental health and existential struggles, it is interesting to see how existential health both might be a buffer against illness and bolster and build health also in these domains. Experiencing mastery and purpose in life, being engaged in something bigger than oneself, and cultivating positive social relationships have been shown to predict better self-rated health, less disability, healthier profiles of biological risk, greater well-being, and better cognitive function in aging adults, even in the context of disability and chronic illness (Ryff et al., 2012). A sense of purpose in life is associated with a reduced risk for all-cause mortality (Cohen et al., 2016). And the value of *not* becoming foreign to negative emotions and being able to relate to a broad diversity in feelings and mental states is also of great importance to mental as well as physical health (Parrott, 2014; Quoidbach et al., 2014). Therefore, health promotion policies also should include such perspectives. However, there is a paradox at play if purpose, meaning, and our capacity to relate to suffering become means with health as a goal. This paradox parallels the one that Yalom (2012) describes when he says that although we need meaning in life, and lack of meaning is deeply disturbing for us, it is something that we are less likely to find the more we deliberately pursue it. Meaning ensues from meaningful activity; meaningfulness is a byproduct of engagement and commitment. Similarly, a good health outcome is more like a bonus we gain from our willingness to face existential challenges; it is not the main reason we engage in them in the first place. Illness will always sooner or later hit us. A concept of existential health might highlight some important possibilities in life: In an existential sense you might die healthy, in a medical sense none of us do.

CONCLUSION

We might become healthier through our existential struggles; meaning is essential for a healthy life. But at the deepest level, we do not face our existential challenges to become healthy. We struggle with them because purpose, meaning, and existential concerns are of the highest value for our lives. In a highly medicalized society, we need to reinvent our language for these existential struggles, and the suffering that always accompanies them. If a concept of "existential health" can make us more aware of the fact that the existential concerns are always part of both health and illness, it is worth developing.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Albarracin, D., Ducousso-Lacaze, A., Cohen, D., Gonon, F., Keller, P.-H., and Minard, M. (2015). There is no cure for existence: on the medicalization of psychological distress. *Ethic. Hum. Psychol. Psychiatry* 17, 149–158.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., and Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychol. Sci. Public Inter.* 4, 1–44.

Baumeister, R. F., Vohs, K. D., Aaker, J. L., and Garbinsky, E. N. (2013). Some key differences between a happy life and a meaningful life. *J. Posit. Psychol.* 8, 505–516.

Bromberg, P. M. (1996). Standing in the spaces: the multiplicity of self and the psychoanalytic relationship. *Contemp. Psychoanal.* 32, 509–535. doi: 10.1080/00107530.1996.10746334

Brownlee, S., Chalkidou, K., Doust, J., Elshaug, A. G., Glasziou, P., Heath, I., et al. (2017). Evidence for overuse of medical services around the world. *Lancet* 390, 156–168. doi: 10.1016/s0140-6736(16)32585-5

Buber, M. (2003). *Between Man and Man*. London: Routledge.

Cohen, R., Bavishi, C., and Rozanski, A. (2016). Purpose in life and its relationship to all-cause mortality and cardiovascular events: a meta-analysis. *Psychosom. Med.* 78, 122–133. doi: 10.1097/PSY.0000000000000274

DeMarinis, V. (2008). The impact of postmodernization on existential health in Sweden: psychology of religion's function in existential public health analysis. *Arch. Psychol. Relig.* 30, 57–74. doi: 10.1163/157361208x316962

Esdaile, A. S., Shah, F., and Binder, P.-E. (2021). *Eksistensiell Helse er Blitt et Nøkkelbegrep Under Pandemien [Existential Health Has Become a Key Concept in the Pandemic]*. Psykologisk.no.

Frankl, V. E. (2004). *Man's Search for Meaning: The Classic Tribute to Hope From the Holocaust*. New York, NY: Random House.

Fugelli, P. (2006). The Zero-vision: potential side effects of communicating health perfection and zero risk. *Patient Educ. Counsel.* 60, 267–271. doi: 10.1016/j.pec.2005.11.002

Gadamer, H.-G. (2018). *The Enigma of Health: The Art of Healing in a Scientific Age*. Hoboken, NJ: John Wiley & Sons.

Gendlin, E. T. (1982). *Focusing*. New York: Bantam.

Gendlin, E. T. (1992). The primacy of the body, not the primacy of perception. *Man and World* 25, 341–353.

Heidegger, M. (1957). *Zein und Zeit [Being and Time]*. Tübingen: Niemeier.

Heidegger, M. (2000). *Kunstverkets Opprinnelse, Med en Innføring av Hans-Georg Gadamer [The Origin of the Work of Art, With an Introduction by Hans-Georg Gadamer]*. Oslo: Pax.

Jaspers, K. (2010). *Philosophy of Existence*. Philadelphia, ON: University of Pennsylvania Press.

Kabat-Zinn, J., and Hanh, T. N. (2009). Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness. New York, NY: Delta Trade.

Kirmayer, L. J. (2000). "Broken narratives: clinical encounters and the poetics of illness experience," in *Narrative and the Cultural Construction of Illness and Healing*, eds C. Mattingly and L. C. Garro (Berkeley,CA: University of California Press), 153–180.

Kissane, D. W. (2012). The relief of existential suffering. *Arch. Int. Med.* 172, 1501–1505. doi: 10.1001/archinternmed.2012.3633

Kornfield, J. (2017). *No Time Like the Present: Finding Freedom, Love, and Joy Right Where You Are*. New York, NY: Simon and Schuster.

Madsen, O. J. (2014). *The Therapeutic Turn: How Psychology Altered Western Culture*. Abingdon: Routledge.

May, R. (1999). *Freedom and Destiny*. New York, NY: WW Norton & Company.

McAdams, D. P., and Bowman, P. J. (2001). "Narrating life's turning points: redemption and contamination," in *Turns in the Road: Narrative Studies of Lives in Transition*, eds D. P., Josselson, R., and Lieblich, A. Washington, DC: American Psychological Association, 3–34.

Melder, C. (2012). The epidemiology of lost meaning: a study in the psychology of religion and existential public health. *Scrip. Instit. Donner. Aboen.* 24, 237–258. doi: 10.30674/scripta.67417

Merleau-Ponty, M. (1962). *Phenomenology of Perception*. London: Routledge.

Miller, R. B. (2005). Suffering in psychology: the demoralization of psychotherapeutic practice. *J. Psychother. Integrat.* 15, 299–336.

Miller, R. B. (2015). *Not so Abnormal Psychology: A Pragmatic View of Mental Illness*. Washington, DC: American Psychological Association.

Norcross, J. C. (2011). *Psychotherapy Relationships That Work: Evidence-Based Responsiveness*. Oxford: Oxford University Press.

Parrott, W. G. (2014). *The Positive Side of Negative Emotions*. New York, NY: Guilford Publications.

Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., and Norton, M. I. (2014). Emodiversity and the emotional ecosystem. *J. Exp. Psychol. Gen.* 143:2057. doi: 10.1037/a0038025

Ratcliffe, M. (2009). Existential feeling and psychopathology. *Philos. Psychiatry Psychol.* 16, 179–194.

Ryff, C., Friedman, E., Fuller-Rowell, T., Love, G., Miyamoto, Y., Morozink, J., et al. (2012). Varieties of resilience in MIDUS. *Soc. Pers. Psychol. Comp.* 6, 792–806. doi: 10.1111/j.1751-9004.2012.00462.x

Sandler, J. (1960). The background of safety. *Int. J. Psycho-Anal.* 41, 352–356.

Schneider, K. J. (1999). Suffering and its ambiguities. *Psychother. Pat.* 11, 109–114. doi: 10.4324/9781315786360-8

Siegel, D. J. (2010). *Mindsight: The New Science of Personal Transformation*. London: Bantam.

Sigurdson, O. (2016). Existential Health. Philosophical and historical perspectives. *LIR J.* 6, 8–26.

Sigurdson, O. (2019). "Only vulnerable creatures suffer: on suffering, embodiment and existential health", in *Phenomenology of the Broken Body*, eds E. Dahl, C. Falke, and T. E. Eriksen (New York, NY: Routledge), 87–100.

Sirois, F. M. (2016). *"Perfectionism and health behaviors: a self-regulation resource perspective," in Perfectionism, Health, and Well-Being*, eds F. M., Sirois and D. S. Molnar. (Berlin: Springer), 45–67.

Solomon, S., Greenberg, J., and Pyszczynski, T. (2015). *The Worm at the Core: On the Role of Death in Life*. New York, NY: Random House.

Suissa, A. J. (2008). Addiction to cosmetic surgery: representations and medicalization of the body. *Int. J. Ment. Health Addict.* 6, 619–630. doi: 10.1007/s11469-008-9164-2

Svenaeus, F. (2019). A defense of the phenomenological account of health and illness. *J. Med. Philos. Forum Bioeth. Philos. Med.* 44, 459–478. doi: 10.1093/jmp/jhz013

Sweeny, K., and Dooley, M. D. (2017). The surprising upsides of worry. *Soc. Pers. Psychol. Comp.* 11:e12311. doi: 10.1111/spc3.12311

Tillich, P. (2000). *The Courage to Be*. London: Yale University Press.

Vaillant, G. (2008). *Spiritual Evolution: A Scientific Defense of Faith*. New York, NY: Harmony.

Whitehead, P. M. (2019). *Existential Health Psychology: The Blind-Spot in Healthcare*. Berlin: Springer.

Winnicott, D. W. (1954). Mind and its relation to the psyche-soma. *Br. J. Med. Psychol.* 27, 201–209. doi: 10.1111/j.2044-8341.1954.tb00864.x

Winnicott, D. W. (1965). Communicating and not communicating leading to a certain opposites. in *The Maturational Process and the Facilitating Environment*. New York, NY: International University Press, 179–192.

Winnicott, D. W. (1971). *Playing and Reality*. London: Burns & Oates.

Winnicott, D. W. (2014). *Through Pediatrics to Psychoanalysis: Collected Papers*. Abingdon: Routledge.

Yalom, I. D. (1980). Existential psychotherapy. 1980. New York, NY: Basic Books.

Yalom, I. D. (2012). *Love's Executioner: & Other Tales of Psychotherapy*. New York, NY: Basic Books.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Binder. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Psychology

HYPOTHESIS AND THEORY
published: 26 May 2022
doi: 10.3389/fpsyg.2022.861584

What Is a *White Epistemology* in Psychological Science? A Critical Race-Theoretical Analysis

*Thomas Teo**

Department of Psychology, York University, Toronto, ON, Canada

Critical race theory guides the analysis of the *nature* of a *white epistemology* in psychological science, the consequences for the study of race, and how scientific racism has been possible in the pursuit of knowledge. The article argues that race has not only been misused in the politics of psychology but misappropriated because of the *logic* of psychological science. The epistemic process is divided into four components to argue that naïve empiricist approaches in psychology, centered on scientific method, prevent an intricate understanding of race. Reasons for privileging method in psychology and the consequences of a *white epistemology* are discussed, including a narrow epistemic horizon and an inability to account for the temporality and contextuality of psychological phenomena. Ignorance, failure, or unwillingness to account for epistemic complexity when studying race are identified as problems. Questions about who benefits from narrow epistemologies are answered and suggestions for a broader practice of knowledge and education are provided.

Keywords: critical race theory, philosophy of science, *epistemology*, knowledge, race, racism, methodologism

OPEN ACCESS

Edited by:
Isabel Menezes,
University of Porto, Portugal

Reviewed by:
Marco Aurelio Prado,
Federal University of Minas Gerais, Brazil
Bolaji Balogun,
The University of Sheffield, United Kingdom

****Correspondence:***
Thomas Teo
tteo@yorku.ca

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology,
a section of the journal Frontiers in Psychology

Received: *24 January 2022*
Accepted: *21 April 2022*
Published: *26 May 2022*

Citation:
Teo T (2022) What Is a White Epistemology in Psychological Science? A Critical Race-Theoretical Analysis. Front. Psychol. 13:861584.
doi: 10.3389/fpsyg.2022.861584

The American Psychological Association (APA) released in October of 2021 an apology that acknowledges the role of APA and the discipline of psychology in "promoting, perpetuating, and failing to challenge racism" (American Psychological Association [APA], 2021a) and a resolution that aims at "dismantling systemic racism" (American Psychological Association [APA], 2021b). The apology acknowledges psychology's history of racism, including its participation in eugenics and the maintenance of racial hierarchies, while calling for perspectives based on human rights, anti-racist approaches in all institutions, and the support for minority psychologists. While "white epistemologies" are identified as general sources of racism in psychology, without being described, the following article intends to address the question of what could be considered a *white epistemology* or in which ways the "logic of research" as the existing practice of psychological science has contributed to the problem. The issue, so goes the argument, is not just the politics and power but the *logic of psychological research*.

Acknowledging that something went wrong in psychology when dealing with race, and even more significantly, understanding the complicity and/or ignorance of (American) psychology with regard to racism in its research and practices, must be complemented by a consideration of the ways in which psychological science, as it has been taught and embodied in the practices of psychologists, has contributed to the problem. Beyond bad intentions, personal biases, or ideologically motivated psychologists, what is "happening" within psychological science itself? To answer the question of how psychological science and its logic have subsidized racism in the discipline and to argue that racism is a meaningful consequence of the logic of traditional research, I will draw on ideas from *Critical Race Theory* (CRT). Focusing on the logic of research, I intend to advance the analysis beyond problems of sampling, reporting, reviewing, and disseminating research (see Buchanan et al., 2021).

CRITICAL RACE THEORY FOR PSYCHOLOGICAL SCIENCE

Because of the confusion that the project of CRT evokes, it is necessary to discuss its meanings. There exist at least three different "language games" when it comes to the term *Critical Race Theory*: (a) CRT refers to an approach within legal studies that has been developed since the 1970s and 1980s; (b) CRT has been applied and extended, explicitly or implicitly, to other academic projects in the humanities, social sciences, and education; and (c) CRT is politically misinterpreted to suggest that *white* people are inherently bad or oppressive, a narrative that has gained political currency, particularly in the United States (see e.g., Hargis and Walker, 2021). This article will draw mostly on (b), which is supported by empirical evidence and able to guide further analyses.

CRT was developed by legal scholars, including Bell (1976) and Freeman (1978), as a framework for studying how race and racism play out in law and the legal system (for overviews, see Delgado and Stefancic, 1993; Crenshaw et al., 1995). An important innovation was made by the critical race theorist Crenshaw (1989), who emphasized the intersectionality of race, gender, class, and other social characteristics in legal and other social contexts (for psychology, see Rosenthal, 2016). Outside legal theory, critical ideas on race have been developed, sometimes with reference to CRT, in education (e.g., Dixson et al., 2017), sociology (e.g., Omi and Winant, 1994), historiography (Pascoe, 1996), as well as psychology (Salter and Adams, 2013; Fine and Cross, 2016; Salter et al., 2018), and other disciplines. For philosophy, Mills (1997) developed the argument that America is based on a racial contract of *white* supremacy, not only politically but also epistemologically.

Theorizing *white epistemology*, this argument draws on core tenets of CRT (for psychology see Salter and Adams, 2013; Delgado and Stefancic, 2017): (a) Although the meaning of race and racism are socially and historically constituted ("racism without races"), the effects of those constructions and racializations are tangible. The process of racialization and the social constitution of race have been shown by anthropological and historical research (e.g., Hannaford, 1996; Yudell, 2014). Thus, *Whiteness* in this article is not understood as a racial category. Similar to Mills's (2007) argument about white ignorance, a *white epistemology* is not confined to *white* people. Indeed, one could use concomitantly the terms dominant, hegemonic, traditional, or Euro-American indigenous epistemology. The article rejects the idea that historically developed race divisions should be treated as "natural kind" categories in research, reified through results of empirically found differences (Teo, 2018).

A *white epistemology* is not inherently a problem, no more than any other indigenous epistemology, but historically has become problematic when it embodies hegemony and relates to the study of *other races*. As CRT points out (b) racism is a common practice, enacted by states, governments, and institutions (for instance, in Northern America), that reflects the economic, social, political, or cultural interests of dominant groups in society (Feagin, 2006). The argument presented here draws on knowledge that "white" people have benefitted from research on race (see Gould, 1996) and analyzes how the logic of research is tilted toward the needs and interests of dominant racialized groups in society (for *Whiteness* see also Lipsitz, 2006). The term *white epistemology* is justified when knowledge benefits one ethno-cultural group to the detriment of another.

Following the previous argument, CRT proposes that (c) race and racism are embedded in structures (e.g., American society), institutions (e.g., the legal system, education, the labor market) and life-worlds that lead to unequal outcomes for racialized people. These outcomes emerge independently of the intentions, attitudes, or behaviors of identifiable individuals. Racism is systemic (Elias and Feagin, 2016) and science is not excluded from this process if one considers science to be part of society. This article develops the idea that a *white epistemology* is embedded in the logic of psychological science. Arguably the *psychological humanities* (see Teo, 2017) have a different logic of research (Gadamer, 1960/1997) and would require a separate analysis (that is not provided here). For instance, one could identify a *white epistemology* within an ignorance (Mills, 2007) that presumes European history to be central and the standard against which all other historiographies be measured, an assumption and practice challenged by postcolonial historians (e.g., Chakrabarty, 2000).

CRT suggests that (d) the voices, perspectives, and first-person experiences of racialized people should be included in research (e.g., Collins, 1991). The following argument agrees that neglecting the voices of racialized people leads to ignorance and distortions about their lives and subjectivities. Not including the voice of the *Other* is indeed an important aspect of a *white epistemology* (see also Said, 1978/1979). However, the proposed perspective suggests that voice is necessary but insufficient should the logic of psychological research not be changed. Thus, this argument drawing on CRT investigates the degree to which a *white epistemology* is embedded in the system of psychological science and its consequences for the study of race (including scientific racism).

Arguably, the American Psychological Association [APA] (2021a,b) resolutions are drawing on ideas of CRT when understanding racism as part of an institution and when considering power in the discipline of psychology. Psychologists and researchers interested in psychology have addressed some of these critical topics from historical points of view (Tucker, 1994; Gould, 1996; Jackson and Weidman, 2004; Winston, 2004; Richards, 2012). Psychologists have also looked at the institutional structure of psychology and its manifestation in psychological research, in terms of sampling, problematic generalizations, editorial boards, and ignorance about race (Arnett, 2008; Henrich et al., 2010; Roberts et al., 2020; Buchanan et al., 2021). It is not surprising that some critical psychologists have understood the discipline and practice as a colonial project, with calls for the decolonization of psychology (e.g., Decolonial Psychology Editorial Collective, 2021).

THE LOGIC OF (PSYCHOLOGICAL) SCIENCE

Using a rational reconstruction, the analysis begins with the question of how and in which ways epistemic injustice is embedded in methodic rationality as a standard practice of truth in psychology. The existing logic of psychological science makes it difficult and implausible, if not impossible, to understand how scientific racism (for examples see Winston, 2020a) has been possible (likewise for scientific sexism and scientific classism). To develop the argument that a *white epistemology* is inherent in the logic of psychological science, an analysis of the components of research is required. Drawing on philosophies of science, the practice of knowledge is divided into the *context of discovery, context of justification, context of interpretation,* and *context of translation*.

The distinction between the context of discovery and context of justification goes back to the logical-positivist philosopher of science, Reichenbach (1938), who studied the natural sciences when making his well-known distinction. He argued that researchers intending to reconstruct knowledge should focus on *internal* relations and not on the *external* sources of knowledge; the latter he considered the domain of sociology and psychology. The study of knowledge should focus, according to Reichenbach, on the internal structure of knowledge, on what he called the *context of justification*. The critical rationalist Popper (1935/1992) made a similar distinction between epistemology and psychology to suggest that the study of science should focus on how statements are tested or justified (deductively for Popper) while excluding questions about the psychological sources that led to "discovery." Both Reichenbach and Popper made the argument that scientific reconstructions of science, the practices of knowledge, should focus on the context of justification, which is considered in this argument to be at the core of a white epistemology in psychological science.

Applied to psychological research, those "positivist" prescriptions meant that psychologists should attend to how they justify their knowledge claims, to whether statements have been tested, verified or falsified, and to whether they provide internally valid, reliable, objective, or generalizable statements based on empirical (preferably experimental) studies. Although the logic of research plays out differently in various academic disciplines, the core idea of this approach is that the claim to knowledge should center on methodology in a narrow sense, more aptly expressed as *method*. Scientific method is the path to knowledge and the only path to knowledge. The context of justification excludes questions about the societal, historical, cultural, interpersonal, or personal sources of knowledge and excludes the reasons why researchers are interested in what they are studying. The context of discovery was banned from epistemic debates by leaders of positivism, of naïve empiricism and in psychology. However, understanding the external sources of knowledge is particularly relevant when studying race.

With historical, sociological, and psychological studies of science pioneered by Fleck (1935/1979) and Kuhn (1962), and with the many works in *Science and Technology Studies* (STS) (e.g., Latour and Woolgar, 1979), epistemologists have learned that an assessment of knowledge requires not only an understanding of the *internal* features of a science (i.e., method) but also an understanding of the history, politics, sociology, and psychology of science (*external* features). With Kuhn (1962), one could ask whether the distinction between the two contexts (discovery and justification) is helpful, given that the actual practice of science involves continuous entanglements between the two (see also Barad, 2006). With the participation of prominent psychologists in scientific racism (see Yakushko, 2019; Winston, 2020a), the distinction between contexts makes little analytic sense (see also Tucker, 2002). The positivism dispute in Europe (Adorno et al., 1969) was grounded in the debate on how political and personal interests, social characteristics, power, money, preconceived notions, or worldviews contribute to or even constitute scientific knowledge. Critical theorists have emphasized that the narrow focus on the context of justification, considered central for a *white epistemology* in this article, limits our understanding of knowledge in substantial ways. To fully understand the substance of knowledge, including what has been studied and, more importantly, what has not been studied in the social sciences—and no less in psychology—one needs to move beyond method.

In addition to these two classical contexts, another component in the practice of research needs to be identified, one which cannot be reduced to the context of justification. Because the data (results) produced, using rigorous methods, are not the same as the interpretation of data, interpretation represents another analytic context. Interpretation is particularly important in the social sciences and in psychology where complexity is part of the subject matter's ontology. The context of interpretation is also grounded in debates in the philosophy of science, particularly in the underdetermination thesis (Duhem, 1905/1954; Quine, 1970), according to which theoretical interpretations of results are underdetermined by data.

The discipline of psychology depends significantly on the discussion of results to make epistemic claims (see also Holzkamp, 1964/1981). Certainly, in actual research, the context of interpretation is entangled with other contexts: The interpretation of data depends not only on the data but also on programmatic commitments, historical mentalities, cultural assumptions, and societal ideas; there exists a circularity between theory and data, when theories produce particular sets of data and when data are then interpreted within those theories. For instance, a study within scientific racism produces data within this framework and data are interpreted within the same framework (see the many examples in Tucker, 1994; Gould, 1996; Winston, 2020b).

Finally, one should add a fourth component of research, which itself is the result of historical shifts in the meaning of science as an institution: The context of "translation," or how research results and interpretations are reported, reviewed, disseminated, used, applied, or implemented in the academic and social world. In the academic world, translation with a focus on scientific impact, involves publishing one's results in journals, books, chapters, or reports, or presenting one's results at conferences (exclusion and inclusion come into play). Research findings must be expressed in linguistic terms using the conventions contained, for instance, within publication manuals. Increasingly there are institutional demands to inform, involve, and engage the public when translating knowledge, and scientific contributions may be assessed in terms of their *public* relevance or impact. Granting agencies and university administrators are asking researchers to target not only the academic community, but also the public, communities, or media. Indeed, measurable "impact" (academic, public, or economic) has become a significant criterion for academic success.

Psychological science is not exempt from increasing pressures to articulate the relevance of research to the public and studies in scientific racism have always had a large impact (see also Jackson and Winston, 2021). The increasing importance of the context of translation is not peripheral to the epistemological project of science, although arguably, the myth persists in psychological science that what really counts, epistemologically, is method. For people constructed through research in a negative way (as deficient, inferior, or damaged), the process of knowledge translation is crucial. A *white epistemology* means excluding or not attending to the contexts of discovery, interpretation, and translation, while at the same time ignorance is produced (see also Mills, 2007).

SCIENTIFIC PSYCHOLOGY'S IDENTITY

In psychology, positivism has had an important role as many historians and theoreticians have pointed out (Tolman, 1992; Teo, 2018). Psychology was one of the disciplines that embraced positivism (Winston, 2001) and centered on the context of justification, translated as focusing primarily on methodology and method, and developing extensive, sophisticated, and complex tools for studying psychological phenomena. The focus (critics could call it obsession) on method, seemingly justifying the status of psychology as a real science, has been labeled *methodolatry* (Bakan, 1967), the *methodological imperative* (Danziger, 1985), or *methodologism* (Teo, 2005), all important dimensions of a *white epistemology*. Psychologists justify psychological knowledge as science by claiming the use of the scientific method and maintain that psychological work is scientific because of its use of method. Less discussed in the discipline is *doing justice to the object* (e.g., race) or the claim that primacy should be given to the object and not the method. One can justifiably describe this focus on method epistemologically as naïve empiricism (Teo, 2018) and one can consider it, given its history, a hegemonic or *white epistemology*.

From an epistemic point of view, a science that includes the subjects as objects of knowledge would require reflexivity regarding sources of questions, methods, interpretations, and translations, given the socio-political track records of the discipline and, for instance, the evidence from debates on intelligence and immigration in the 1920s (Gould, 1996) and, more recently, the participation of psychologists and psychological institutions in torture practices (Hoffman et al., 2015). The primacy of method in scientific psychology entails that issues that emerge in the context of discovery are not considered relevant. However, as studies on scientific racism in psychology have demonstrated, the reconstruction of political, social, historical, and economic sources of knowledge would contribute to a better understanding of psychological work (Jackson and Weidman, 2004). Arguably, even the replication crisis (Open Science Collaboration, 2015) cannot be solved with methodology alone but requires an understanding of the nature of the psychological that may be characterized by contextuality and temporality. Let me illustrate through the use of a "thought experiment": If the Nazis had won the war and had come to dominate the rest of the world, and found empirical differences in characteristics between Aryans and segregated Non-Aryans using scientific methods, adding that many of the characteristics have high heritability estimates, then would it be fair to conclude that those differences are natural or a given? A method focused on difference (and on statistically assessing difference) would not be able to address that question or to challenge the theoretical and practical assumptions that formed the basis of the scientific study. A focus on method alone, that is, a white epistemology, would lead to misleading knowledge.

The argument that an assessment of psychological knowledge should include not only the context of justification, but also the contexts of discovery, interpretation, and translation, is not an attempt to narrow but to broaden the meaning of science by reconstructing knowledge in its full complexity. The lack of attention to all contexts limits psychological knowledge in significant ways. For instance, the 5th edition of the American Psychological Association [APA], 2001 Publication Manual argues that psychologists are "free to examine, interpret, and qualify the results, as well as to draw inferences from them" (p. 26). From the perspective of this argument, one should add that although psychologists are free, they also have an epistemic responsibility to provide good and careful interpretations, interpretations that do justice to the topic under investigation. The ability to provide good interpretations of data and results is a skill that needs to be taught, learned, and developed as much as any other epistemic skill in psychology. However, most psychology programs do not offer textbooks or courses on teaching hermeneutic skills in the contexts of interpretation, discovery, and translation. It is assumed that the context of justification (method) solves all other epistemic problems, which cements a training focus on research methods and statistics. Yet, a modern scientific psychology, beyond a *white epistemology*, needs to account for all contexts if knowledge is to remain a priority.

The focus on method seems to justify the scientific status of psychology or enables the discipline to move up in the hierarchy of sciences. However, this focus reproduces epistemic ignorance about what it means to be a science. The outcome in psychology is not a true science but a *hyperscience* that resembles a science (by imitating and developing methods) without a full account of the psychological object (Teo, 2020). Scientific method alone, the use of small or big machines, or the rhetoric of science, can only provide a semblance of science. The obsession with being acknowledged as a science has historical and cultural roots; it became evident when it was beneficial for financial and public reasons to associate academic psychology with the natural sciences and not with history, philosophy, or the humanities (Ward, 2002). Any gaze into the history of psychology supports that argument. For instance, Woodworth (1921), a participant in scientific racism, uses the terms *scientific* and [Frame4] *science* 13 times on the first page of his textbook. The pioneer of American psychology, James (1890), whose own book would be dismissed from a presentist perspective as unscientific, also claimed psychology as a real science (indeed, the *history of psychology* could be a source for discussing each context).

All pioneers of psychology struggled with identifying the subject matter of psychology, that is, with attending to ontic questions in psychology. Yet, from a positivist point of view, theorizing the subject matter of psychology should take backstage to research methods. This focus on method has led to some of the problems discussed in indigenous, cultural, feminist or critical psychologies as well as in theoretical psychology (see Teo, 2018) and also engenders the re-occurring crises debates in the discipline (see also Wieser, 2016). Attempts to transform traditional criteria such as validity into *psychopolitical validity* (Prilleltensky, 2008), or accounting for the degree to which an intervention captures both the psychological and the political, are rare. The apparent increasing usage of qualitative methods in psychology expands the range of phenomena that count as scientific (Gergen et al., 2015). However, qualitative methods, from the perspective of this argument, are not superior and require the same ontic and epistemic discussions as quantitative methods and may have their own *white epistemic* assumptions (see also Chauhan and Sehgal, 2022). To be fair, qualitative researchers emphasize the need to add more epistemic dimensions to the assessment of knowledge, including reflexivity (e.g., Finlay and Gough, 2003).

When it comes to issues of race, and beyond a *white epistemology*, a complex understanding of the scientific process must include not only a concentration on method but must pose questions about the financing of studies, about their intent and purpose, and about the motivation for the interest in racialized group differences. More generally, such an understanding must address when, where, and how the concept of "race" (or intelligence, etc.) and its study emerged; the economic and cultural interests that engendered such studies; and social and political interests (context of discovery). This does not mean ignoring the methodological and methodic shortcomings of such studies (context of justification). An understanding of the scientific process must also include an analysis of the quality of interpretations and the relationship between theory and data and must emphasize that results do not determine interpretations (context of interpretation). Such an analysis may focus particular attention on forms of *epistemological violence* as practices that are executed when academic interpretations of empirical results implicitly or explicitly construct the "Other" as problematic or inferior, even though alternative interpretations exist (Teo, 2008). Finally, a thorough understanding of research on races should include an analysis of explicit or implicit recommendations, concrete applications, and discourses about racialized life for the public (context of translation) (e.g., should different races really be sent to different schools?).

PROBLEMATICS FOR THE STUDY OF RACE IN PSYCHOLOGICAL SCIENCE

Engaging CRT means to understand how scientific racism and racism have been produced, and cannot be avoided in the traditional logic of research on race with its focus on method. Drawing on the discussions above, three problematics emerge for understanding *white epistemology*. The focus on method leads to (a) a *narrow horizon* that excludes discussions and reflections about all components of research that must come into play when race is studied. For a scientific understanding of race or race differences, knowledge from a variety of disciplines, including history, sociology, anthropology, political science, legal studies and the philosophy of science, are required (see also Nelson et al., 2013; Bonam et al., 2019). Excluding these fields may produce limited knowledge and ignorance in scientific psychology. Thus, whenever race is used as a variable in psychological research, a move from the narrow horizon of method to all contexts that are relevant in knowledge-making is necessary.

It is fair to ask whether this narrow horizon means that a *white epistemology* permeates scientific psychology generally, or only when race is included in its studies, or only when working from the perspective of scientific racism. Generally, one can argue that a hegemonic (from a descriptive point of view) or *white* (from a historical point of view) epistemology infuses the whole discipline. Because Western culture frames the discipline, it is reasonable to ask to what degree psychology has ethnic, cultural or colonial biases, not only when it comes to studies that include race, but in basic psychological research that attempts generalizations (Teo and Febbraro, 2003). Whiteness, not understood here as a racial but as a sociological category, has permeated knowledge in psychology, including cognitive psychology and perception research (for examples see Henrich et al., 2010). Thus, it may not be racism in a narrow sense but rather a sense of Western centrality and/or superiority that guides assumptions about generalizability (as a reaction we find attempts to develop and justify an African psychology; see Nwoye, 2015). It is not always a racial, but more recently, a cultural project that assumes the superiority of Western modes of psychological thinking (Teo, 2022). From this perspective it is evident that a *white epistemology* needs to be problematized in the study of race and it becomes toxic in scientific racism.

An analysis of scientific racism should focus on the conditions that have made this project possible. Because the argument applies to psychological science as a system, historical examples would deflect attention away from the academic discipline to individual actors, which would undermine a core tenet developed in CRT. However, to elucidate the point, two examples that corroborate the logic of scientific racism are mentioned. The eugenicist Charles Davenport (1866-1944) illustrates perfectly how the focus on method misses the complexity of the problem. Davenport and Steggerda's (1929) book on "race mixture" appears to be a prime example of pure objective empirical science, with hundreds of tables, figures, and numbers, and using batteries of tests, including psychological ones, coming to the scientific conclusion that "intermingling" would be bad.

In psychology, J. Philippe Rushton's (1943-2012) work can be considered paradigmatic for scientific racism. His empirical studies on "three major races" and claims that certain races are more aggressive, less intelligent, and less law-abiding than others (Rushton, 1995) – using traditional psychological methods and instruments, without discussing the contexts of discovery (e.g., worldview), interpretation, and translation, and not including the voices of the *Other* who are rendered damaged and deficient through his ideas – represent an ideal form of *white epistemology*. It is for this very reason, a lack of understanding of the complexity of the problem of race, that Rushton's papers in *Psychological Reports* have been retracted (Retraction Notice, 2021).

Hiding behind a horizon based on methodologism shows limitations when complexity is reduced to the point of distortion. If one finds, within the logic of scientific psychology, differences between two racialized groups, there is nothing in method itself that prevents one from theorizing, interpreting and concluding that those differences are natural. This logic explains how naïve empiricism was able to produce racist research (and biased research on other social characteristics) that was peer-reviewed and published but is unable to address the socio-historical constitution of those social characteristics. The latter would entail posing questions about interests in scientific racist research or the degree to which scientific racism is a *Weltanschauung* oblivious to counterevidence (Winston, 2020b). Ideological or incompetent interpretations that suggest that within-group heritability estimates can be used to explain between-group differences still need to be addressed (see Tucker, 1994).

Overall, the history of psychology demonstrates that naïve empiricist psychology has not only been unable to prevent racist (or sexist and classist) research but has encouraged it within its logic. Challenges to scientific racism have stemmed from expanding the narrow horizon of method. The call for decolonizing psychology follows from such experiences (Decolonial Psychology Editorial Collective, 2021). Decolonial and anti-racist work (e.g., Jones, 2018) is possible when race is not treated as a natural kind entity, but rather as an entity that has social, cultural, and historical dimensions.

The second issue that has plagued empiricist psychology as a *white epistemology* is (b) the *status quo supporting* role of scientific psychology. If psychology does not specifically include an anti-racist or decolonial position and instead focuses primarily on method, then it is inevitable that research in psychology reproduces the *status quo*. If one lives in a racist society, then empirical (scientific) results will reflect that racist society. *Status quo* research could find that group X has lower scores on intelligence tests than group Y, or lower scores on motivation, achievement-orientation, and so on, without tracing the entanglement of the history of race, racism, intelligence testing, and politics. Such research on *status quo* differences then reinforces racist thinking or the problematization of the group that has been construed as inferior from the beginning (this can even happen with good intentions as textbooks show). In Northern America, *status quo* supporting research reinforces Whiteness, thus assisting the interests and needs of groups that are already dominant in society.

More generally, one could argue that psychology has problems with a broad understanding of temporality. Psychology has methods that account for temporality in longitudinal research or in pre- and post-test studies. Such methods can account for developmental changes in perception, cognition, identity, affects, and so on, over a lifespan. But beyond evolutionary, age-based or situational temporality, psychology needs to account for historical changes and the ways in which history constitutes and shapes mental life (Pettit and Hegarty, 2014). History is full of psychosocial content and to understand that content for subjectivity, psychology must encompass not only the psychological sciences but also the psychological humanities (Teo, 2017). Race and racism cannot be fully understood without an understanding of the history of race and racism in various locations (see also Nelson et al., 2013). The exclusion of such histories produces ignorance about the entanglement of historical, social, and cultural contexts for a seemingly simple variable. The focus on empirically found differences, without accounting for political, economic, or cultural developments, neglects important dimensions of mental life. In short, scientific methods, for all their strengths, do not account for histories of oppression, marginalization, extermination, violence, injustice, and unequal power. Historically constituted realities such as slavery and institutional discrimination have produced different life-worlds for ethnic groups in Northern America.

Historical thinking allows for a vision not only of what is, but also of what is possible. Beyond descriptive questions of what is possible are normative questions of what should be. It is not uncommon in many scientific disciplines, from medicine to climate science, to combine descriptions, predictions, and normative reflections. Yet, good normative reflections require a deep understanding of knowledge in the psychological humanities (e.g., philosophy). Arguably, a broader horizon on psychosocial issues (including race) sets the conditions for the possibility of a fuller discussion of normative issues than a narrow horizon. To be fair, psychology has put forth concepts and methods for future possible developments. Such concepts and methods range from Vygotsky's (1978) *zone of proximal development*, which attempts to assess what is possible for an individual, to *Participatory Action Research* (Fine and Torre, 2021), which was developed in communities to study not only what has been, but also what can be changed and achieved when people act together, and which considers marginalized persons as co-researchers. Yet, there is no place in the context of justification to consider what is possible or what justice could mean when it concerns issues of race.

Temporality needs to be accounted for, as does (c) *contextuality.* Psychological concepts, theories, methods, measures, and practices have a cultural dimension and they reflect that culture, which could be labeled an indigenous challenge to the discipline (see also Sundararajan et al., 2020). To be sure, scientific psychology has developed cross-cultural tools to investigate differences or similarities, but culture also includes the cultural constitution of scientific concepts, theories and methods. It would be strange to argue that there are cultural differences, but that they do not apply to psychological science itself. Indeed, the idea that knowledge should focus on the context of justification and that psychology should focus on method is itself a cultural (and historical) product (thus, a *white epistemology*). Although scientific methods may find empirical differences between group X and group Y, the reasons for the differences or the meaning of these differences cannot be answered within the traditional scientific method. Interpreting differences without accounting for temporality or contextuality can lead to a banality of violence.

It is evident that scientific psychology has a hegemonic European and American history and culture (Walsh et al., 2014) and that some Western countries have been reconstructed as dominant sources of academic psychology, although there have been differences between German, British or French models of doing psychology (Danziger, 1990). Arguably, scientific psychology reflects those cultures and traditions, and psychology has an indigenous *white* dimension in those very contexts, which has been distributed to the rest of the world (where they have been accepted, modified, or rejected). Thus, it is legitimate to ask to what degree Western psychology sees differences, competencies, or performances not only from the perspective of the West, but from a perspective that reflects the intellectual, economic or social interests of groups or individuals in those cultures. More radically, following CRT, one can ask whether scientific psychology is perpetuating white supremacy or the degree to which psychology represents the need for the cultural supremacy of the West—particularly the cultural supremacy of the United States, which has had such a large impact on the field of psychology (see also Liu et al., 2019).

An analysis of cultural, economic or social interests, sometimes shared by individual researchers, and issues concerning cultural supremacy, needs to include knowledge of cultures, ethnicities, races or other groups, and cannot be conducted within the context of justification. To make it clearer: Method *prevents* the asking of such questions if they are not posed as empirically testable hypotheses. Yet, such an analysis is required, given the track record of psychology with scientific racism. Equally so, issues of cultural supremacy cannot be solved within the scientific method but must rely on knowledge from the humanities and other social sciences, including critical race theories. Processes of *Othering* or of *Inferiorizing the Other*, regardless of intent, requires historical, political, and cultural knowledge that cannot be found within scientific psychology alone (although research on racism in scientific psychology can assist this project).

Some have made the argument that the process of *inferiorizing* in psychology is no different from that in other disciplines. However, other disciplines do not make the same kinds of inferiorizing interpretations as are made within psychology. For example, if technical advances in physics are made by a group from a particular geographical region, this does not mean that groups from other geographical regions are inferior, that is, it is not suggested within the discipline of physics that other groups are inferior because of a lack of similar technological advances. That type of interpretation is already the work of scientific racism or culture-supremacy. In psychology, an analysis of the mental life of another culture from the perspective of a dominant culture has an inherent cultural problem, even without invoking empirical differences, ranking and quantification. An understanding of mental life, as some of the pioneers of the discipline understood (Dilthey, 1957), requires an understanding of customs, traditions, and the social system. It also requires reflexivity because of the looping effects and the *human kind* qualities of psychological concepts (see Hacking, 1994) and the entanglement of the object with the subject of research. It requires reflexivity on the culture and subculture of the researcher, which may draw on the psychology, sociology, and history of science (see O'Doherty et al., 2019). For an assessment of the quality of psychological research, the inclusion of all contexts of research is required, arguably more so than in other disciplines.

CONCLUSION

The logic of traditional psychological research, often combined (but not always) with a *Weltanschauung* (Winston, 2020b), supports the *status quo* and is unable to address the power of racism in the discipline and profession. This logic is grounded, explicitly or implicitly, in the focus on the context of justification that conceals the importance of other contexts. In elevating method, traditional psychological science is unqualified to address scientific racism. In not moving outside method, psychological science may participate in systemic racism in the conduct of research on races. In contrast to the traditional approach of psychological science (as a *white epistemology*), the study of race in psychology needs to incorporate a discussion of the contexts of discovery, interpretation, and translation. This involves the integration of knowledge from history, culture, politics, sociology, philosophy, economics, anthropology and other disciplines (e.g., genetics) about race. To be clear: The argument advanced here does not suggest that assessing issues within the context of justification, such as methodic quality in research, is irrelevant. Rather, the issue is that method is insufficient when it comes to issues of race. I submit that most researchers recognize intuitively the need for epistemic complexity, and that topics such as race require the psychological sciences as well as the knowledge produced by the psychological humanities, including the critical, postcolonial, and anti-racist ones, but this is hardly expressed in textbooks on research practices.

Doing justice to the topic of race in research requires not only changes in research practices but also changes in education and training. The teaching of psychology must move beyond an exclusive focus on method and must encourage critical thinking not only as a methodological virtue but in relation to all contexts of research, as well as in regard to reflexivity about one's own assumptions in science and about race. This requires teaching a model of psychological science in which method does not have primacy but which also brings social, political and financial interests, and the power of/in interpretation and translation to the foreground. Education on the history of science, the history of psychology, the philosophy and sociology of science, and hermeneutics, among other humanities, would provide a general and broad knowledge base. Yet, these are exactly the courses that have been considered superfluous in scientific psychology.

Scholars have recommended the decolonization of psychology and an active anti-racist stance to overcome the racism of psychology and to avoid invoking superiority or inferiority in discussions of race. Critical race theory in a narrow or broad sense remains an important resource for thinking about race and its implications when studying psychological phenomena. Reflections in this area as well as the willingness to learn about uncomfortable histories and actualities in the discipline and profession would provide the conditions for the possibility of a more responsible psychological science. In my experience of teaching the historical and theoretical foundations of psychology, the large majority of students, even when not following debates about scientific racism, are aware that method alone misses important dimensions of mental life. There is also an openness among students to learn about methods that attempt to do more justice to complex issues, that do not work with variables, but that rather seek to capture a problem in its complexity; one such example is a *circuits of dispossession* methodology that accounts for contextuality and temporality (e.g., Fine and Ruglis, 2009).

Given the complexity and the accumulated knowledge about race and racism, it may also be important to encourage epistemic modesty in education when making knowledge claims about race. This includes an analysis of the neoliberalization of academia, which revokes the idea of modesty when celebrating the marketing of research. Outrageous claims about race and violent interpretations are accompanied by academic citations and public debates (Jackson and Winston, 2021). The reward structure of academia adds to the probability of epistemic grandiosity rather than modesty (Teo, 2019). From an epistemic point of view, science means doing justice to the problem of race and requires a broad horizon of many streams of knowledge, and because the development of this broad horizon is a long-term project, modesty seems appropriate from an epistemic point of view.

It is fair to ask how this argument accounts for alternative and disrupting projects (e.g., challenging racism) that have existed in the history of psychology. My answer is that those projects were possible not because but despite the existing logic of research. While scientific racism has always claimed science to bear witness, some of the most important interventions in psychology that challenged scientific racism have been based on historical and theoretical reconstructions (Chorover, 1979; Howitt and Owusu-Bempah, 1994; Gould, 1996; Richards, 2012) that used cultural, social and philosophical knowledge and material. This is not deny that methodical critiques remain significant when analyzing scientific racism. The move from studying race to studying racism (see also Samelson, 1978), for instance, the study of color-blind racism (Neville et al., 2016), has been made possible because of an expansion of the epistemic project from a narrow context to the broadening of sources, interpretations, and applications of psychology.

The argument is not a call for censorship. It is a call for a better interrogation of knowledge and race in psychological science, for a broadening of our horizons, and for a more comprehensive science of psychology. The call for epistemic responsibility when it comes to race and racism is not a demand for limiting research but rather an invitation to extend the boundaries of research. Encouraging accountability regarding the knowledge that exists on race is not about suppression, nor is CRT suspect when it serves to address such issues. For psychology, theories, concepts, and methods are needed that show the entanglement of the societal, interpersonal, and personal, and that allow us to understand the complexity of the subjective, including racialized subjectivities, the subjectivities of supremacy, and the subjectivities of researchers who conduct studies on race. Theorizing in psychology remains an important task, in relating empirical research to theories, developing new and drawing on existing theories, and connecting the general with the particular. CRT is one condition for the possibility of a psychology that moves the discipline from making racialized groups of people into problems to aiding in solving problems that racialized people encounter when living their everyday lives.

AUTHOR CONTRIBUTIONS

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

FUNDING

This research has been supported by the Social Sciences and Humanities Research Council of Canada (No. 435-2017-1035).

ACKNOWLEDGMENTS

I would like to thank Angela R. Febbraro and Mike Pettit for their helpful suggestions and comments.

REFERENCES

Adorno, T. W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., and Popper, K. R. (1969). *Der Positivismusstreit in der Deutschen Soziologie [The Positivism Dispute in German Sociology]*. Munich: Luchterhand.

American Psychological Association [APA] (2001). *Publication Manual of the American Psychological Association*, 5th Edn. Washington, DC: American Psychological Association.

American Psychological Association [APA] (2021a). Apology to People of Color for APA's Role in Promoting, Perpetuating, and Failing to Challenge Racism, Racial Discrimination, and Human Hierarchy in U.S. [Resolution Adopted by the APA Council of Representatives on October 29, 2021]. Available online at: https://www.apa.org/about/policy/racism-apology (accessed December 30, 2021).

American Psychological Association [APA] (2021b). *Role of Psychology and APA in Dismantling Systemic Racism Against People of Color in U.S. [Resolution adopted by the APA Council of Representatives on October 29, 2021]*. Available online at: https://www.apa.org/about/policy/dismantling-systemic-racism (accessed December 30, 2021).

Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *Am. Psychol.* 63, 602–614. doi: 10.1037/0003-066X.63.7.602

Bakan, D. (1967). *On Method: Toward a Reconstruction of Psychological Investigation*. San Francisco, CA: Jossey-Bass.

Barad, K. M. (2006). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.

Bell, D. A. (1976). Serving two masters: integration ideals and client interests in school desegregation litigation. *Yale Law J.* 85, 470–516. doi: 10.2307/795339

Bonam, C. M., Nair Das, V., Coleman, B. R., and Salter, P. (2019). Ignoring history, denying racism: mounting evidence for the Marley hypothesis and epistemologies of ignorance. *Soc. Psychol. Pers. Sci.* 10, 257–265. doi: 10.1177/1948550617751583

Buchanan, N. C. T., Perez, M., Prinstein, M. J., and Thurston, I. B. (2021). Upending racism in psychological science: strategies to change how science is conducted, reported, reviewed, and disseminated. *Am. Psychol.* 76, 1097–1112. doi: 10.1037/amp0000905

Chakrabarty, D. (2000). *Provincializing Europe: Postcolonial Thought and Historical Difference*. Princeton, NJ: Princeton University Press.

Chauhan, A., and Sehgal, S. (2022). Interrogating paradigmatic commitments of focus group methodology: an invitation to context-sensitive qualitative research methods. *Qual. Psychol.* [Epub ahead of print]. doi: 10.1037/qup0000227

Chorover, S. L. (1979). *From Genesis to Genocide: The Meaning of Human Nature and the Power of Behavior Control*. Cambridge, MA: MIT Press.

Collins, P. H. (1991). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. London: Routledge.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Univ. Chic. Leg. Forum* 1989, 139–167. doi: 10.4324/9781003199113-14

Crenshaw, K., Gotanda, N., Peller, G., and Thomas, K. (eds) (1995). *Critical Race Theory: The Key Writings that Formed the Movement*. New York, NY: New Press.

Danziger, K. (1985). The methodological imperative in psychology. *Philos. Soc. Sci.* 15, 1–13. doi: 10.1177/004839318501500101

Danziger, K. (1990). *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge: Cambridge University Press.

Davenport, C. B., and Steggerda, M. (1929). *Race Crossing in Jamaica*. Washington, DC: Carnegie Institute of Washington.

Decolonial Psychology Editorial Collective (2021). General psychology otherwise: a decolonial articulation. *Rev. Gen. Psychol.* 25, 339–353. doi: 10.1177/10892680211048177

Delgado, R., and Stefancic, J. (1993). Critical race theory: an annotated bibliography. *Va Law Rev.* 79, 461–516. doi: 10.2307/1073418

Delgado, R., and Stefancic, J. (2017). *Critical Race Theory: An Introduction*, 3rd Edn. New York, NY: New York University Press.

Dilthey, W. (1957). *Die Geistige Welt: Einleitung in die Philosophie des Lebens (Gesammelte Schriften V. Band) [The Mental World: Introduction to the Philosophy of Life (Collected Writings, Vol. 5)]*. Stuttgart: Teubner.

Dixson, A. D., Anderson, C. R., and Donner, J. K. (2017). *Critical Race Theory in Education: All God's Children Got a Song*, 2nd Edn. New York, NY: Routledge.

Duhem, P. (1905/1954). *The Aim and Structure of Physical Theory* (P. P. Wiener, trans.). Princeton, NJ: Princeton University Press.

Elias, S., and Feagin, J. R. (2016). *Racial Theories in Social Science: A Systemic Racism Critique*. New York, NY: Routledge.

Feagin, J. R. (2006). *Systemic Racism: A Theory of Oppression*. New York, NY: Routledge.

Fine, M., and Cross, W. E. Jr. (2016). "Critical race, psychology, and social policy: refusing damage, cataloging oppression, and documenting desire," in *The Cost of Racism for People of Color: Contextualizing Experiences of Discrimination*, eds A. N. Alvarez, C. T. H. Liang, and H. A. Neville (Washington, DC: American Psychological Association), 273–294. doi: 10.1037/14852-013

Fine, M., and Ruglis, J. (2009). Circuits and consequences of dispossession: the racialized realignment of the public sphere for U.S. youth. *Transform. Anthropol.* 17, 20–33. doi: 10.1111/j.1548-7466.2009.01037.x

Fine, M., and Torre, M. E. (2021). *Critical Participatory Action Research*. Washington, DC: American Psychological Association.

Finlay, L., and Gough, B. (2003). *Reflexivity: A Practical Guide for Researchers in Health and Social Sciences*. Malden, MA: Blackwell Science.

Fleck, L. (1935/1979). *The Genesis and Development of a Scientific Fact*. Chicago, IL: University of Chicago Press.

Freeman, A. D. (1978). Legitimizing racial discrimination through antidiscrimination law: a critical review of Supreme Court doctrine. *Minn. Law Rev.* 62, 1049–1119.

Gadamer, H.-G. (1960/1997). *Truth and Method* (J. Weinsheimer and D. G. Marshall, trans.). London: Continuum.

Gergen, K. J., Josselson, R., and Freeman, M. (2015). The promises of qualitative inquiry. *Am. Psychol.* 70, 1–9. doi: 10.1037/a0038597

Gould, S. J. (1996). *The Mismeasure of Man (Revised and Expanded)*. New York, NY: Norton.

Hacking, I. (1994). "The looping effects of human kinds," in *Causal Cognition: A Multi-Disciplinary Approach*, eds D. Sperber, D. Premack, and A. J. Premack (Oxford: Clarendon Press), 351–382. doi: 10.1093/acprof:oso/9780198524021.003.0012

Hannaford, I. (1996). *Race: The History of an Idea in the West*. Baltimore, MD: Johns Hopkins University Press.

Hargis, C., and Walker, A. (2021). *Fox News' Critical race Theory Obsession. Media Matters for America*. Available online at: https://www.mediamatters.org/fox-news/fox-news-critical-race-theory-obsession (accessed May 7, 2021).

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83. doi: 10.1017/S0140525X0999152X

Hoffman, D. H., Carter, D. J., Lopez, C. R. V., Benzmiller, H. L., Guo, A. X., Latifi, S. Y., et al. (2015). *Report to the Special Committee of the Board of Directors of the American Psychological Association: Independent review relating to APA Ethics Guidelines, National Security Interrogations, and Torture*. Sidley Austin LLP. Available online at: http://www.apa.org/independent-review/APA-FINAL-Report-7.2.15.pdf (accessed December 30, 2021).

Holzkamp, K. (1964/1981). *Theorie und Experiment in der Psychologie: Eine Grundlagenkritische Untersuchung (Zweite, um ein Nachwort Erweiterte Auflage) [Theory and Experiment in Psychology: A Study Critical of its Foundations*, 2nd Edn. Berlin: De Gruyter.

Howitt, D., and Owusu-Bempah, J. (1994). *The Racism of Psychology: Time for Change*. New York, NY: Harvester Wheatsheaf.

Jackson, J. P., and Weidman, N. M. (2004). *Race, Racism, and Science: Social Impact and Interaction*. Santa Barbara, CA: Abc-Clio.

Jackson, J. P., and Winston, A. S. (2021). The mythical taboo on race and intelligence. *Rev. Gen. Psychol.* 25, 3–26. doi: 10.1177/1089268020953622

James, W. (1890). *The Principles of Psychology*, Vol. 2. New York, NY: Holt.

Jones, C. P. (2018). Toward the science and practice of anti-racism: launching a national campaign against racism. *Ethn. Dis.* 28, 231–234. doi: 10.18865/ed.28.s1.231

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

Latour, B., and Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills, CA: Sage.

Lipsitz, G. (2006). *The Possessive Investment in Whiteness: How White People Profit from Identity Politics (rev. and expanded ed.)*. Philadelphia, PA: Temple University Press.

Liu, W. M., Liu, R. Z., Garrison, Y. L., Kim, J. Y. C., Chan, L., Ho, Y. C. S., et al. (2019). Racial trauma, microaggressions, and becoming racially innocuous: the role of acculturation and White supremacist ideology. *Am. Psychol.* 74, 143–155. doi: 10.1037/amp0000368

Mills, C. W. (1997). *The Racial Contract*. Ithaca, NY: Cornell University Press.

Mills, C. W. (2007). "White ignorance," in *Race and Epistemologies of Ignorance*, eds S. Sullivan and N. Tuana (Albany, NY: State University of New York Press), 13–38.

Nelson, J. C., Adams, G., and Salter, P. S. (2013). The Marley hypothesis: denial of racism reflects ignorance of history. *Psychol. Sci.* 24, 213–218. doi: 10.1177/0956797612451466

Neville, H. A., Gallardo, M. E., and Sue, D. W. (eds) (2016). *The Myth of Racial Color Blindness: Manifestations, Dynamics, and Impact*. Washington, DC: The American Psychological Association.

Nwoye, A. (2015). What is African psychology the psychology of? *Theory Psychol.* 25, 96–116. doi: 10.1177/0959354314565116

O'Doherty, K. C., Osbeck, L. M., Schraube, E., and Yen, J. (2019). *Psychological Studies of Science and Technology*. Cham: Springer International Publishing.

Omi, M., and Winant, H. (1994). *Racial Formation in the United States: From the 1960s to the 1990s*, 2nd Edn. New York, NY: Routledge.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349, aac4716. doi: 10.1126/science.aac4716

Pascoe, P. (1996). Miscegenation law, court cases, and ideologies of "race" in twentieth-century America. *J. Am. Hist.* 83, 44–69. doi: 10.2307/2945474

Pettit, M., and Hegarty, P. (2014). "Psychology and sexuality in historical time," in *APA Handbook of Sexuality and Psychology (Vol. 1: Person-based Approaches)*, eds D. L. Tolman, L. M. Diamond, J. A. Bauermeister, W. H. George, J. G. Pfaus, and L. M. Ward (Washington, DC: American Psychological Association), 63–78. doi: 10.1037/14193-003

Popper, K. R. (1935/1992). *The Logic of Scientific Discovery*. London: Routledge.

Prilleltensky, I. (2008). The role of power in wellness, oppression, and liberation: the promise of psychopolitical validity. *J. Commun. Psychol.* 36, 116–136. doi: 10.1002/jcop.20225

Quine, W. V. (1970). On the reasons for indeterminacy of translation. *J. Philos.* 67, 178–183. doi: 10.2307/2023887

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Chicago, IL: The University of Chicago Press.

Retraction Notice (2021). Retraction notice. *Psychol. Rep.* doi: 10.1177/00332941211042507 [Epub ahead of print].

Richards, G. (2012). *"Race", Racism and Psychology: Towards a Reflexive History*, 2nd Edn. London: Routledge.

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., and Mortenson, E. (2020). Racial inequality in psychological research: trends of the past and recommendations for the future. *Perspect. Psychol. Sci.* 15, 1295–1309. doi: 10.1177/1745691620927709

Rosenthal, L. (2016). Incorporating intersectionality into psychology: an opportunity to promote social justice and equity. *Am. Psychol.* 71, 474–485. doi: 10.1037/a0040323

Rushton, J. P. (1995). *Race, Evolution, and Behavior*. Brunswick, NJ: Transaction.

Said, E. W. (1978/1979). *Orientalism*. New York, NY: Random House.

Salter, P., and Adams, G. (2013). Toward a critical race psychology. *Soc. Pers. Psychol. Compass* 7, 781–793. doi: 10.1111/spc3.12068

Salter, P. S., Adams, G., and Perez, M. J. (2018). Racism in the structure of everyday worlds: a cultural-psychological perspective. *Curr. Dir. Psychol. Sci.* 27, 150–155. doi: 10.1177/0963721417724239

Samelson, F. (1978). From "race psychology" to "studies in prejudice": some observations on the thematic reversal in social psychology. *J. Hist. Behav. Sci.* 14, 265–278. doi: 10.1002/1520-6696(197807)14:3&lt;265::aid-jhbs2300140313&gt;3.0.co;2-p

Sundararajan, L., Hwang, K.-K., and Yeh, K.-H (eds) (2020). *Gloabl Psychology from Indigenous Perspectives: Visions Inspired by K.S. Yang*. Cham: Palgrave Macmillan.

Teo, T. (2005). *The Critique of Psychology: From Kant to Postcolonial Theory*. Boston, MA: Springer. doi: 10.1007/b107225

Teo, T. (2008). From speculation to epistemological violence in psychology: a critical-hermeneutic reconstruction. *Theory Psychol.* 18, 47–67. doi: 10.1177/0959354307086922

Teo, T. (2017). From psychological science to the psychological humanities: building a general theory of subjectivity. *Rev. Gen. Psychol.* 21, 281–291. doi: 10.1037/gpr0000132

Teo, T. (2018). *Outline of Theoretical Psychology: Critical Investigations*. London: Palgrave Macmillan, doi: 10.1057/978-1-137-59651-2

Teo, T. (2019). "Academic subjectivity, idols, and the vicissitudes of virtues in science: epistemic modesty versus epistemic grandiosity," in *Psychological Studies of Science and Technology*, eds K. O'Doherty, L. Osbeck, E. Schraube, and J. Yen (London: Palgrave Macmillan), 31–48. doi: 10.1007/978-3-030-25308-0_2

Teo, T. (2020). Theorizing in psychology: from the critique of a *hyper-science* to conceptualizing subjectivity. *Theory Psychol.* 30, 759–767. doi: 10.1177/0959354320930271

Teo, T. (2022). "Culture-supremacy: expressions, sources, and resistance to a psychology of motivated ignorance," in *Research in the Social Scientific Study of Religion: Lesser Heard Voices in Studies of Religion*, Vol. 32, eds R. W. Hood and S. Cheruvallil-Contractor (Leiden: Brill), 323–340.

Teo, T., and Febbraro, A. (2003). Ethnocentrism as a form of intuition in psychology. *Theory Psychol.* 13, 673–694. doi: 10.1177/09593543030135009

Tolman, C. W. (ed.) (1992). *Positivism in Psychology: Historical and Contemporary Problems*. New York, NY: Springer-Verlag.

Tucker, W. H. (1994). *The Science and Politics of Racial Research*. Urbana, IL: University of Illinois Press.

Tucker, W. H. (2002). *The Funding of Scientific Racism: Wickliffe Draper and the Pioneer Fund*. Urbana, IL: University of Illinois Press.

Vygotsky's, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: Harvard University Press.

Walsh, R., Teo, T., and Baydala, A. (2014). *A Critical History and Philosophy of Psychology: Diversity of Context, Thought, and Practice*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139046831

Ward, S. C. (2002). *Modernizing the Mind: Psychological Knowledge and the Remaking of Society*. London: Praeger.

Wieser, M. (2016). Psychology's "crisis" and the need for reflection. A plea for modesty in psychological theorizing. *Integr. Psychol. Behav. Sci.* 50, 359–367. doi: 10.1007/s12124-016-9343-9

Winston, A. (2001). "Cause into function: Ernst Mach and the reconstruction of explanation in psychology," in *The Transformation of Psychology: Influences of 19th-Century Philosophy, Technology, and Natural Science*, eds C. D. Green, M. Shore, and T. Teo (Washington, DC: American Psychological Association), 107–131. doi: 10.1037/10416-006

Winston, A. S. (2020a). "Scientific racism and North American psychology," in *Oxford Research Encyclopedia of Psychology*, ed. O. Braddick (Oxford: Oxford University Press).

Winston, A. S. (2020b). Why mainstream research will not end scientific racism in psychology. *Theory Psychol.* 30, 425–430. doi: 10.1177/0959354320925176

Winston, A. S. (ed.) (2004). *Defining Difference: Race and Racism in the History of Psychology*. Washington, DC: American Psychological Association.

Woodworth, R. S. (1921). *Psychology: A Study of Mental Life*. New York, NY: Holt.

Yakushko, O. (2019). Eugenics and its evolution in the history of western psychology: a critical archival review. *Psychother. Polit. Int.* 17:e1495. doi: 10.1002/ppi.1495

Yudell, M. (2014). *Race Unmasked: Biology and Race in the Twentieth Century*. New York, NY: Columbia University Press.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Teo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

CONCEPTUAL ANALYSIS
published: 06 June 2022
doi: 10.3389/fpsyg.2022.925359

Disturbance of Ecological Self and Impairment of Affordance Perception

Nam-Gyoon Kim[1]* and *Judith A. Effken*[2]

[1] Department of Psychology, Keimyung University, Daegu, South Korea, [2] College of Nursing, University of Arizona, Tucson, AZ, United States

Affordance, a radical concept James Gibson introduced in the 1970s, remains controversial today. Defined as environmental properties taken with reference to an animal's anatomy and action capabilities, affordances are opportunities for action the environment offers. By perceiving affordances, organisms hold meaningful relationships with their surroundings. Affordance is not just a theoretical concept but, as the embodiment of meanings and values, has serious psychological implications. We contend that the lack of these meanings and values underlies the irrational behavior seen in patients with self disorders such as schizophrenia. We reason that it is by perceiving affordances that individuals keep in touch with their surroundings and stay mentally healthy. Using contrapositive reasoning, the reverse could also be true. That is, when individuals experience difficulty maintaining meaningful relations with their surroundings and suffer from mental health problems, we might anticipate that their affordance detection systems are impaired. In two studies conducted in our laboratory, patients with schizophrenia and Alzheimer's disease were shown to have impaired capacity to perceive affordances, a result qualifying as contra-positive evidence corroborating the affordance concept. In addition, our results provide support for accepting contra-positive evidence as a complementary tool to positive evidence for empirically validating concepts such as affordance and meaning.

Keywords: affordance, disturbance of minimal self, schizophrenia, Alzheimer's disease, contra-positive evidence, ecological self, empirical investigation of meaning

OPEN ACCESS

Edited by:
Anna M. Borghi,
Sapienza University of Rome, Italy

Reviewed by:
Zakaria Djebbara,
Aalborg University, Denmark
Julia J. C. Blau,
Central Connecticut State University, United States

****Correspondence:***
Nam-Gyoon Kim
nk70@kmu.ac.kr

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *21 April 2022*
Accepted: *17 May 2022*
Published: *06 June 2022*

Citation:
Kim N-G and Effken JA (2022) Disturbance of Ecological Self and Impairment of Affordance Perception. Front. Psychol. 13:925359. doi: 10.3389/fpsyg.2022.925359

INTRODUCTION

Of the many ideas put forth by the American perceptual psychologist James Gibson, *affordance*, without question, is the most radical (Gibson, 1977, 1979, 1982). The conventional account of how an animal interacts with the surrounding environment begins with physical energies impinging on sensory receptors. When stimulated, the receptors transduce the energies into neural signals that spread across cortical and subcortical regions while undergoing several stages of enrichment. Such elaboration processes are mandated because the impoverished sensory input is devoid of meaning (having been produced by the meaningless physical entities comprising the environment) and therefore cannot represent the surroundings adequately. However, animals routinely interact with the environment (i.e., with objects, places, events, and other animals) in meaningful ways.

Gibson rejected a dualistic stance that separates an (objective) physical world devoid of meanings from (subjective) mental states replete with meanings. Instead, Gibson envisioned that animal and environment are reciprocally conjoined, thus forming an inseparable pair. To portray the reciprocal relationship of animal and environment, Gibson defined the properties of the environment, not in units of pounds, feet, and seconds, but in units of the animal's body dimensions and action capabilities, i.e., affordances (Warren, 1984; Mark, 1987). With its properties defined as affordances, the environment now abounds with behavior-relevant properties that offer opportunities for animals to act. Moreover, with environmental properties expressed in animal-referential terms, its descriptors now are compatible with those of the animal, enabling the animal's contact with the environment to be direct, epistemic, and meaningful. As Gibson (1979) put it, "If what we perceived were the entities of physics and mathematics, meanings would have to be imposed on them. But if what we perceive are the entities of environmental science, their meanings can be discovered" (p. 33).

Suppose that a human animal has a specific goal in mind. It actively searches its surroundings for an affordance that can help it reach its goal. When a relevant affordance is identified, the human animal uses that affordance to implement and fine-tune its actions until its goal is realized. Subserved by an affordance, this encounter with the environment yields a unique experience. Each experience, however insignificant or trivial, involves the combined effort of the perception-action system. Significantly, whether it is reaching out and grasping a mug or pulling a chair and sitting on it, successful accomplishment of an intended action is a manifestation of autonomy and control, assuring one's sense of agency. This may be why each experience serving as the embodiment of meaning motivates the animal to continue interacting with its surroundings.

What might transpire if the human animal's capacity to access affordances is somehow compromised? No longer able to appreciate affordances or maintain meaningful relationships with others or with its surroundings, the consequences for the animal would be dire. This is likely the case with some of the most devastating mental disorders, particularly those disorders phenomenological psychopathologists suspect to arise from disturbance of the minimal self (i.e., self-disorder). As their ability to register affordances decreases, these human animals become increasingly deprived of meanings and values, eventually developing severe mental health symptoms. Based on this reasoning, we propose and demonstrate the use of an indirect way to validate the concept of affordance. We further propose utilizing patients with mental disorders suspected to be caused by disturbance of minimal self as a testing ground to validate the concept of affordance, and at the same time, provide an empirical investigation of meaning.

In contemporary cognitive science, "people disappear and are replaced by symbolic constructs and manipulations analogous to those of computer programs" (Reed, 1996, p. 3). Gibson and his followers have rejected this approach by restoring the experiences and activities of persons and animals to psychological reality. A similar sentiment has been raised by Varela et al. (1991) who noted that "cognitive science has had virtually nothing to say about what it means to be human in everyday, lived situations" (p. xv). The philosophical perspective put forth by Varela and his followers is known as enactivism [Varela et al. (1991), Noë (2004), and de Haan (2020), for review]. Since enactivism and ecological psychology both denounce mental representation to account for cognition, several members of ecological camp have considered whether an integrated conceptual framework is possible. Thus far, all have found these efforts futile [see Flament-Fultot et al. (2016), Heft (2020), and Read and Szokolszky (2020)]. The irreconcilability between the two perspectives lies in the way sensation is conceived. Further discussion of these philosophical positions and their differences is beyond the scope of the present study so we suggest that the reader refer to the cited references for further details.

In the following, we further delineate the concepts of affordance and ecological self and then describe our own psychophysical studies assessing affordance perception capacity.

AFFORDANCES

A hiker went for a long hike in Central Park in New York City. After several hours, the hiker encountered a horizontal, flat, extended, and rigid surface at approximately knee height. Whether that surface is a park bench, a tree stump, or a swing, it offers an opportunity for the hiker to sit and rest. However, flat, extended, and rigid surfaces provide places to sit only for those individuals whose lower leg length corresponds roughly to the height of the seats.

Clearly, a sit-on-able surface for an adult is different from that for a 2-year old, but both offer sit-on-able affordances. Thus, a sit-on-able affordance exists, irrespective of an individual's age and/or physical makeup. In this sense, affordances are objective properties. However, a sit-on-able surface is uniquely tailored, not only for an individual's specific body dimensions, but also for that individual's specific needs and circumstances. For example, for someone with a painful, swollen hip, sit-on-able is not an affordance a hard park bench offers. Thus, affordances are also subjective. As Gibson (1979) noted, "an affordance is neither an objective property nor a subjective property; or it is both. . . (that is, it) cuts across the dichotomy of subjective—objective" (p. 129).

When defined in reference to an individual's action capabilities, the environment offers many opportunities for the individual to act. As was the case with the hiker described above, an individual perceives an affordance that would fulfill his needs at a given moment by detecting the information specifying that affordance. Visually, the ambient light structure at an observation point is uniquely determined by the composition and layout of the surrounding surfaces and is specific to the affordances those surfaces offer. For example, rigid surfaces engender patterns that differ from those of elastic surfaces (von Fieandt and Gibson, 1959). Although the number of surfaces comprising our surroundings is infinite and of many types, the number of dimensions along which surfaces can vary is finite, thus limiting the number of optical invariants to which an organism must attune.

To exploit the available environmental information about a sit-on-able surface, our hiker had to seek it actively. In the words of Gibson (1979), "We must perceive in order to move, but we must also move in order to perceive" (p. 223). In the case of vision, this involves not only using the eyes, but "the eyes in the head on the shoulders of a body that gets about" (Gibson, 1979, p. 222). The information specifying an affordance must be actively detected by our hiker, but the hiker's action is guided by perceptual information. As the hiker moves, the ambient energy distribution is uniquely transformed in accordance with the changes in the environmental layout and the displacements of the observation point. This transforming energy pattern at a moving point of observation (i.e., optic flow) is specific both to the environmental layout and the animal's movements that engendered it (Gibson, 1979; Warren, 1998, 2006, 2021). Our hiker's forward movement structured the optic flow such that all optical elements radiated from a single point (i.e., the focus of expansion) corresponding to the hiker's own movement direction. By regulating the direction of movement coincident with the focus of expansion, our hiker was able to reach the intended target, then sit down and rest.

Although the structured light is ambient about an observation point, our hiker could sample only a portion of the optic array due to a limited field of view [see Figure 7.1 in Gibson (1979), p. 113]. That field of view, if portrayed as an oval window, contained various optical structures, some of which corresponded to the hiker's body parts (e.g., orbits of the eyes, nose, upper lip, cheeks, and limbs). As the hiker moved (e.g., turning from side to side), those optical structures corresponding to body parts transformed. Because the optical transformation was produced by the hiker's movement, the transformation patterns were specific to those movements. For Gibson, perception of the environment and perception of the self were inseparable, always occurring together (Reed, 1996). As Gibson (1979) remarked, "One perceives the environment and co-perceives oneself" (p. 126).

In the optic array, information specific to the environment is called extero-specific and information specific to the observer is called proprio-specific. Since, Sherrington (1906), it has been thought that self-perception is conveyed by information from mechanoreceptors in the muscles, tendons, and joints. For Gibson, however, awareness of self can also be gained visually (e.g., as *visual kinesthesis).*

"Vision is *kinesthetic* in that it registers movements of the body just as much as does the muscle-joint-skin system and the inner ear system. . . . Visual kinesthesis goes along with muscular kinesthesis. The doctrine that vision is exteroceptive, that it obtains 'external' information only, is simply false. Vision obtains information about *both* the environment and the self" (Gibson, 1979, p. 183, italics original).

Gibson (1977, 1979, 1982) defined a set of affordances as an *ecological niche*. All living organisms are equipped with the capacity to perceive affordances, which enables them to exploit the myriad of affordances the surrounding environment offers. However, if one's capacity to perceive affordances has been compromised, the consequences are likely to be devastating because the environment has ceased to be meaningful. This appears to be the case in patients with certain clinical and mental disorders [e.g., schizophrenia, post-traumatic stress disorder (PTSD), and Alzheimer's disease (AD)]. A common feature of these disorders is a disturbance of what might be called, as Gallagher (2000, p. 15) did: the "basic, immediate, or primitive 'something' that we are willing to call a self."

ECOLOGICAL SELF

As noted earlier, optic flow is determined by facts about the environment and facts about the observer. Optic flow structure, therefore, can be decomposed into two components, one determined by the environment and the other by the observer. When an observer moves, the entire flow field is disturbed (a global transformation); but when an object in the environment moves, it perturbs the flow field locally (a local transformation) (Fajen and Kim, 2002). As an observer moves, producing a global transformation in the optic flow, the observer is immediately aware of causing this transformation. Neisser (1988) referred to such self-specification in optic flow as "ecological self." As an active agent in the immediate environment, the ecological self "perceives themselves, among other things: where they are, how they are moving, what they are doing, and what they might do, whether a given action is their own or not" (Neisser, 1993, p. 4).

Ecological self can be understood as what phenomenological philosophers call "ipseity" (ipse in Latin meaning "self" or "itself"), also called minimal self, core self, or proto self. The minimal self is characterized by two separable modalities of pre-reflective and primitive subjective awareness, that is, a sense of ownership (awareness of being the source of phenomenal experiences) and a sense of agency (awareness of being the agent executing one's own actions) (Gallagher, 2000; Sass and Parnas, 2003; Zahavi, 2005; Stanghellini, 2009; Fuchs, 2010; Parnas and Sass, 2010; Nelson et al., 2014).

Given its diverse (e.g., positive, negative, and disorganized) mental symptoms, schizophrenia is arguably the most debilitating, yet the most perplexing, of all mental disorders (Arango and Carpenter, 2011). Currently, a schizophrenia diagnosis is based on criteria defined in the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorder (DSM-5) and the World Health Organization's International Statistical Classification of Diseases and Related Health Problems (ICD-10), as well as on structured interviews. The present diagnostic system aims to identify the underlying cognitive and neurobiological processes for each symptom or symptom group comprising schizophrenia's psychopathology (Persons, 1986; Cahill and Frith, 1996). Although recent progress in neuroscience and molecular genetics has furthered our understanding of this disability [see Weinberger and Harrison (2011), for a review], the exact cause of this disease remains elusive (Wong and Van Tol, 2003; Insel, 2010; Jablensky, 2010).

Recently, proponents of phenomenological psychiatry and philosophy have underscored the subjective experience of the patient as a valuable tool for gaining an in-depth understanding of the disorder. These researchers suggest that the disparate psychopathological symptoms of schizophrenia may actually be manifestations of a single phenomenological core: *disturbance of ipseity*. The sources of this self-distortion are thought to be two mutually interdependent processes—hyper-reflexivity and diminished self-presence. Hyper-reflexivity refers to an intensified self-consciousness directing the patient's focal attention to internal feelings; while diminished self-presence refers to a weakening sense of self existence (Sass and Parnas, 2003).

Individuals' awareness of their own thoughts, actions, perceptions, feelings, or pain operates at a pre-reflective (i.e., direct, immediate, implicit, or non-conceptual) level. With alterations of self, patients may feel their presence in this world diminish and their grip on the world slip away, or feel that they are falling under the control of an alien. Their perceived presence disintegrates, receding into the background; and the boundary separating their perceived selves from others vanishes (Parnas, 2000, 2003, 2012; Sass, 2003a,b, 2014; Sass and Parnas, 2003; Cermolacce et al., 2007; Parnas and Sass, 2010; Raballo et al., 2011; Nelson et al., 2014). Simultaneously, an opposite process is underway. The patients' own bodies begin to feel strange and unfamiliar, inviting their explicit attention. As their self-monitoring increases, the surrounding world no longer draws their attention. Gradually their focus of attention shifts inward to reflect on their own mental activity. As self-directed reflection intensifies, aspects of their awareness may separate or detach as if they were external objects. As their selves become more alienated from their bodies and aspects of their own feelings, their actions and expressions no longer feel natural and may result in delusions of alien influence (Fuchs, 2009, 2013).

Phenomenological psychopathologists conceptualize schizophrenia as a disorder triggered by disturbance of minimal self that alienates self from the body and from the world, as depicted above. For Gibson (1979), affordance links the environment and the observer as an inseparable dual: "the awareness of the world and one's complementary relations to the world are not separable" (p. 141). If affordance is what connects the self and the world (i.e., animal and the environment), then the disruption of affordances, for example, due to an impairment in the capacity to perceive them would be expected to separate the self and the world. With impaired affordance perception capacity, patients would increasingly fail to register affordances. As their affordance capacity further deteriorates, their surroundings, once replete with values and meaning, ultimately becomes a barren field. Entrapped in meaningless surroundings, the patients can no longer maintain meaningful relations with environmental entities so retreat from social interactions and other activities, gradually disconnecting from reality until they take on the characteristics of a soulless body or a disembodied spirit (Stanghellini, 2009).

PSYCHOPHYSICAL INVESTIGATION OF AFFORDANCE PERCEPTION CAPACITY

Human artifacts are designed with specific functions in mind. However, these artifacts often provide more than one affordance owing to their multiple properties (e.g., shape, size, material composition, etc.). When a specific tool is not available, we often use other household items to carry out functions beyond those for which they were originally designed if the items provide affordances subserving our intended goal. For example, lacking a screwdriver, we might instead use a coin to drive a screw. Flatness and rigidity, two properties of a coin, provide a "drive-a-screw-able" affordance.

A set of diverse objects, each with a different primary affordance, can offer the same secondary affordance. For example, a bowl or, a jam jar, can be used to collect water from the faucet but so can a shoe or a safety helmet. Kim et al. (2022) used a secondary affordance to assess schizophrenia patients' capacity to perceive affordances. Because of the documented decline across a wide range of cognitive domains in schizophrenia, the experiment was administered using a Go/No-Go protocol, a well-established procedure to assess decision-making in a wide variety of contexts, but simple enough to facilitate patients' cooperation and completion of the task.

For the experiment, three pairs of mutually exclusive affordances were used: (a) *scoop-with/pierce-with*; (b) *pour-in-able/stretchable*; (c) *cut-able-with/mop-up-with*. Each affordance was represented by three objects sharing the same secondary affordance. Thus, six objects comprise one affordance pair (O_{aff1}, O_{aff1}) wherein O_{aff1} had the first affordance (e.g., *scoop-with*) but not the second (e.g., *pierce-with*); O_{aff2} had only the second affordance, but not the first. In each pair of affordances, one served as the target signal and the other as the distractor.

Schizophrenia patients were less accurate and slower than controls. However, when assessed for their capacity to detect the object's physical properties (color, shape, material composition) in a control experiment, schizophrenia patients performed as accurately as controls and faster than they had in the affordance perception task. Based on these findings, the authors concluded that affordance perception capacity is likely impaired in patients with schizophrenia.

APRAXIA OF TOOL USE

Apraxia is a neurological disorder characterized by a marked impairment in performing skilled movements in response to a verbal command, despite intact sensory and motor abilities and comprehension of the task. Although apraxia is a predominant symptom in patients with left brain damage (LBD) after a cerebral vascular accident, it is also one of the core features of dementia and is included in diagnostic guidelines for AD. Recently, a growing number of studies have begun to explore the impact of AD on consciousness (Weiler et al., 2016; Bajic et al., 2021; Bomilcar et al., 2021). We now know that neuropsychiatric symptoms (e.g., depression, delusions and hallucination, agitation, and aggression) are common in AD. Whereas schizophrenia has garnered intense interest from phenomenology, little is known about the impact of AD on minimal self. Of interest, in this regard, is Pazzaglia and Galli (2014), who proposed that apraxia be considered as a disturbed sense of agency. Given that sense of agency constitutes one of the two defining features of minimal self, if (as Pazzaglia and Galli contend) sense of agency is disturbed in patients with apraxia, it is equally likely that patients' perceptual capacity for affordances is disturbed as well.

One significant aspect of apraxia is that it affects a person's ability to use commonly available tools or adapt other objects in the surrounding environment as tools to solve a given problem. Clearly, impaired ability to use tools will limit an individual's functional capacity. Currently, two dominant competing hypotheses (manipulation-based and reasoning-based) try to account for variations in tool use behavior. The manipulation-based account relies on semantic memory, which is accumulated through prior sensorimotor experiences with a particular tool. In this view, an individual, when using a familiar tool, retrieves information from stored sensorimotor experiences (manipulation knowledge) about the tool's purpose, its target object, and the typical movement associated with the tool. The reasoning-based account views the problem a potential tool user faces as an instance of problem-solving in which the individual uses mechanical knowledge to reason about the structural properties of tools and their action targets to solve the current problem.

Based on their literature review of studies examining tool use disorders in LBD patients, Baumard et al. (2014) concluded that failure of mechanical knowledge was the cause of the tool use deficit in LBD patients. Lesourd et al. (2016) investigated whether impaired tool use in AD is the same as in LBD. When given mechanical problem-solving tasks, AD patients (despite difficulties) engaged in strategies like trial-and-error to solve problems, a pattern not seen in LBD patients. The authors concluded that impairment of mechanical knowledge does not underlie tool use deficit in AD, but left open the question of the source of tool use impairment in these patients.

Proponents of the manipulation-based account of tool use suggest that, in cases involving novel tools, affordances may facilitate their applications. Kim et al. (2022) repeated the experiments conducted in Kim and Kim (2017) with four groups, AD, mild cognitive impairment (MCI), Parkinson's disease (PD), and elderly controls (EC). The AD group performed poorest, followed by MCI, PD, and EC, in that order. EC and PD groups performed comparably. AD patients responded randomly to stimuli. MCI patients' performance did not differ significantly from PD, EC, or AD groups, suggesting only a slight degradation in performance. In a control experiment in which participants were asked to report the physical properties of the same objects, all four groups performed reliably. These results provide preliminary evidence that affordance perception capacity is impaired in patients with AD and MCI.

DISCUSSION

When we open our eyes upon waking, many things arise in our fields of view, for example, a spouse, a pet, familiar furniture, and the layout of the room, with the additional smell of coffee aroma and the sound of the coffee maker gurgling from the kitchen. These things revitalize our mind and body. The orbits of the eyes, nose, cheek, upper lip, and limbs projected to the same location in our fields of view assure us that we are alive and (at least somewhat) ready to begin the day's activities. For phenomenological philosophers, what is conveyed through our fields of view on opening our eyes is "mine-ness," (i.e., recognition that we are the agent and the owner of our own actions, experiences, thoughts, and feelings). This is ipseity, i.e., the minimal self.

Upon opening our eyes, eye level immediately scales the surrounding layout and surfaces in units of "eye height" (Warren, 1984; Mark, 1987; Warren and Whang, 1987; Wraga, 1999). Partitioned in terms of eye-height, our surroundings reveal their various affordances (i.e., sit-on-able places, grasp-able objects, pass-able openings, drink-able liquids, edible foods, view-able displays, pet-able pets, etc.) for us to use as needed.

Phenomenological psychiatrists and philosophers have offered the ipseity disturbance hypothesis to account for the symptoms of schizophrenia, the most debilitating and most perplexing, of all mental disorders. The ipseity disturbance hypothesis assumes that two interdependent processes (hyper-reflexivity and diminished self-presence) disturb the minimal self, which in turn disturbs awareness of reality (one's "grip" or "hold" on the world), eventually causing the patient to become disembodied and alienated from the surrounding world (Sass and Parnas, 2003; Fuchs, 2009).

To date, considerable effort has been devoted to further elucidate the phenomenology of self-disorders. Also drawing interest among phenomenological researchers is the search for the biological substrates of minimal self-disturbance in schizophrenia (Kyselo, 2016; Nelson and Sass, 2017; Nelson and Sass, 2017). Although their efforts to unpack seemingly incomprehensible utterances of patients are commendable, many puzzling questions remain, one of which is the connection between the disturbance of minimal self and the two processes underpinning the manifested symptoms of schizophrenia. Why does self-disturbance trigger these two processes? Conversely, why is the minimal self so susceptible to these two processes? Curiously, these issues have rarely been discussed by phenomenological researchers.

Setting these issues aside, we characterize what philosophers call the minimal self as the *ecological self*. Note that there are no clear criteria to define the minimal self except for some vague intuitive feeling of "a basic, immediate, or primitive "something" that we are willing to call a self" (Gallagher, 2000, p. 15). The ecological self, on the other hand, is defined based on an invariant pattern in the optical structure specific to it, that is, a global transformation of the optic array. Indeed, the specification of ecological self as a global transformation of optic flow was cleverly demonstrated by David Lee's now classic "swinging room" research. Lee (Lee and Aronson, 1974; Lee and Lishman, 1975) constructed a room with a fixed floor, but with walls and ceiling that can be swung back and forth. When placed in this room and the walls moved, observers swayed in accordance with the optic flow pattern engendered by the moving room. This swaying occurred despite the fact that balance information (i.e., interoceptive information) provided by the inner ear and the receptors in the muscles and joints signaled that the observers' postures were stationary.

Neisser (1993) credited the ecological self as the first form of self to develop in early infancy. To that extent, we construe ecological self as equivalent to minimal self. Thus, it is ecological self that is likely altered in schizophrenia. As in the case with the question pertaining to the particular symptoms manifested in schizophrenia, we have yet to determine how an altered ecological self would manifest. One plausible rationale might be that the locus of the ecological self may coincide with the locus of the observation point to which optical angles subtended by the observer's body parts project. When a patient starts to experience alienation of the self from the body, it may be that the ecological self has separated (at least in part) from the observation point. As the ecological self gradually drifts away from the observation point, the geometric relationship between environmental objects and their corresponding optical angles is no longer preserved. Such disruption would have an immediate impact, distorting scale factors such as eye height, which would perturb the proprioceptive information specifying the ecological self.

For Gibson, perception of the environment and perception of self are co-implicative. Assuming Gibson is correct, if patients' perceptions of self are disturbed, then, relatedly, their attunement to exteroceptive information (facts about the external environment) is likely to be disturbed, as well. With their perceptual capacity disturbed, these patients can no longer tune into the information specifying affordances. Consequently, their surroundings that once abounded with values and meanings become, for them, a barren field of value-neutral physical objects. We hypothesized that inability to appreciate the affordances comprising their surroundings might underlie the various symptoms manifested by schizophrenia patients. The results of Kim and Kim (2017) corroborated that hypothesis. Schizophrenia patients performed poorly when asked to identify unintended functions of human-made artifacts, but retained their capacity to identify physical properties (e.g., color, shape, and material composition) of the same objects.

Apraxia is defined as the inability to perform skilled actions despite intact sensory and motor abilities. To further explore the claim that apraxia is a manifestation of lost sense of agency (the key factor defining minimal self), Kim et al. (2022) administered the same task performed previously by schizophrenia patients to patients with AD. Patients with MCI, PD, and EC also participated in the study. The AD group performed poorest, followed by MCI, then PD and EC (differences for the latter two were not statistically different). The AD group responded randomly to stimuli, their performance not differing from chance. However, when asked to report the physical properties of the same objects, all four groups performed reliably.

Affordance remains a highly controversial concept. As a theoretical concept, most discussions of affordance have involved clarification of its ontological status. Warren's (1984) seminal stair riser research led initial efforts to validate this concept empirically. Yet, as a concept founded on the principle of mutuality binding the reciprocal pairs of proprioception-exteroception, perception-action, animal-environment, and subjective-objective, designing a testing ground for an empirical validation of affordance comprehensive enough to encompass these aspects of dualities has been a challenge. We explored an indirect way to validate affordance by seeking contra-positive evidence.

In logic, the contra-positive of a conditional statement (if P, then Q) is formed by negating both antecedent and consequent and reversing them (if not Q, then not P) (where P stands for the antecedent and Q stands for the consequent). Thus, "If A, then B" is a direct proof, whereas, "If not B, then not A" is a proof by contra-positive, and these two are logically equivalent. Whereas the concept of affordance can be proven by direct evidence, we show how it can also be proven by contra-positive evidence.

As underscored above, affordances enable individuals to hold meaningful relationships with their surroundings. Those individuals whose capacity to tune into the information specifying affordances is somehow disturbed would be unable to detect affordances. Any dysfunction in affordance perception capacity would block access to meanings and values for these individuals, leaving only value-neutral physical objects. Imprisoned in meaningless surroundings, these individuals would be unable to keep in touch with their immediate environment. Eventually they become alienated, withdrawing from their surroundings and from other individuals. We reason that if one is capable of perceiving affordances, one can hold meaningful relationships with one's surroundings and stay mentally healthy. By contra-positive logic, we can also reason conversely that, if one suffers from severe mental health symptoms and even withdraws from others and from one's surroundings, one's affordance perception capacity must be dysfunctional, depriving values and meanings from the individual, thus preventing the individual to keep in touch with one's surroundings. We have described two studies conducted in our laboratory in which patients with schizophrenia (Kim and Kim, 2017) and AD (Kim et al., 2022) performed an affordance perception task. In both studies, the patients with schizophrenia performed poorly in comparison to healthy elderly controls and patients with other neurodegenerative disorders (e.g., PD). These results demonstrate a deficiency in affordance perception capacity that qualifies as contra-positive evidence for the concept of affordance.

So far, our discussion has focused on those patients with mental disorders whose capacity to perceive affordances has been severely impaired. In contrast, de Haan et al. (2013) studied people with Obsessive-Compulsive Disorder (OCD) whose symptoms were so severe that the only treatment available to them was deep brain stimulation (DBS). Much to everyone's delight, the impact of DBS was remarkable, producing a profound change in patients' experience of being in the world. Taking an eclectic position between enactivism and ecological psychology, the authors attempted to explain the phenomenological changes these patients experienced after DBS as a change in the field of relevant affordances. For example, if depicted as a 3D bar graph, the field of relevant affordances for normal individuals included bars of various heights and colors. For depressed patients, however, the field was shown as gray bars of the same short height to underscore how inconspicuous their surroundings were to them. For OCD patients, the field was depicted as a few tall and brightly colored bars to highlight the voracious consumption of their attention and obsessions.

Whereas de Haan et al. (2013) focused on characterizing patients' experiences after DBS implantation in terms of the configuration of fields of affordances, ultimately the utilization of affordances hinges on the individual's ability to register them. If the patient's capacity to perceive affordances (i.e., the capacity to tune into the information specifying affordances) is impaired, despite how salient a particular affordance might be, the individual would not be able to realize it. Not having assessed DBS patients' affordance perception capacity, as we did with patients with schizophrenia in Kim and Kim (2017) or AD in Kim et al. (2022), we cannot be definitive in our conclusions. However, we suspect that these patients would fit well into the same conceptual framework we used to explain the performance of the patients who participated in our studies. Specifically, we suggest that OCD likely impaired these patients' capacity to perceive affordances, thus entrapping them in an environment with few affordances. DBS then restored the patients' affordance perception capacity, enabling them to rejoice in the abundant affordances comprising their surroundings. For now, we remain curious as to whether the perceptual capacity to detect affordances of these OCD patients would have been similarly impaired as the patients with schizophrenia or AD were in our two studies (Kim and Kim, 2017; Kim et al., 2022).

In the two studies referred to above, we observed a substantial deficit in the capacity to perceive affordances for patients with schizophrenia and AD. With their perceptual capacity for detecting affordances impaired, these patients may find it difficult to keep in touch with their surroundings. Nevertheless, as demonstrated in these studies, these patients are still capable of detecting physical properties, suggesting that they should be able to manage contact with the surroundings to some extent. However, the kind of contact they can manage with the world becomes what Heft refers to as the "second-order mode of knowing." To engage in this mode of knowing, Heft (2003) contends, one has to "step outside of the ongoing flow of immediate perception-action awareness by reflecting on the things of the environment; that is, [one has to] shift the necessarily selective character of [one's] attentional focus from experiencing the immediate flow of events to experiencing the experience and, in doing so, isolate particular portions of immediate experience, holding them in awareness for analysis, categorization, or other second-order or indirect acts of cognition" (p. 151). Heft goes on to describe this mode: "accompanying these acts of reflexivity is a comparative heightening of awareness, as entities in experience are momentarily lifted out of the perceptual flow for closer scrutiny." Heft's description reminds us of hyper-reflexivity, one of the two characteristic processes disturbing patients with schizophrenia.

Taken together, these patients may be able to experience physical objects, but only as neutral things devoid of any psychological values. Not being able to relate to these objects (i.e., not being able to perceive affordances), they are not "drawn toward them or repelled by them for any intrinsic qualities they possess" (Heft, 2003, p. 151). Thus, these patients appear as if they are detached from the world.

CONCLUSION

The standard account of (visual) perception starts with the light reflected from the surface of an object. Upon reaching sensory receptors, the light is converted to neural signals which then travel through various areas in the brain where they are embellished with the aid of the information stored in the memory. As purely mechanized responses to meaningless input signals, meaning is absent until semantic memory intervenes.

Meaning motivates animals. Consequently, animals are attracted to affordances that convey the meanings emerging from the objects with which they interact. Affordance is not just an important concept of a particular psychology theory but, as embodiments of meanings and values, it entails serious psychological implications. If an animal's capacity to apprehend affordances is disabled, that animal would be deprived of objects' meanings. Bereft of the motivation affordances offer, the animal may no longer engage with its surroundings. Eventually the animal would be alienated, both from itself and from the world, becoming a disembodied self or spiritless body (Fuchs, 2009, 2013).

To date, the search for direct evidence for affordance's validity has been conducted primarily by assessing the capacity of healthy participants to perceive affordances [Warren (2021), for review].[1] However, given the psychological values of affordance, a contra-positive statement can also qualify as valid. As noted earlier, a conditional statement (if P, then Q) can be formulated such that: if an individual perceives affordances (P), the individual comes in relationship with values and meanings, thereby maintaining meaningful relationships with the environment (Q). A contra-positive statement (if not Q, then not P) would be: If an individual is deprived of values and meanings, eventually experiencing severe mental suffering (not Q), the individual must have been unable to perceive affordances (not P).

This contra-positive statement appears to be true for patients with mental and clinical disorders that are presumed to be caused by disturbance of ipseity or self-disorder. In two studies conducted in our laboratory, we found that the capacity to perceive affordances was severely impaired in schizophrenia patients and AD patients, an existence proof corroborating affordance. Recently, some authors have contended that the ipseity disturbance (or self-disorder) model can extend to other mental disorders such as PTSD (Ataria and Horovitz, 2021), depersonalization disorder, and panic disorder (Sass et al., 2018). It is important to determine whether a similar decline in affordance perception capacity can be observed in these populations as in schizophrenics (Kim and Kim, 2017) and AD (Kim et al., 2022).

Affordances, when perceived, are used to regulate the action needed to attain an intended goal. Action, in turn, fine-tunes the perceptual system to be more sensitive to invariants specifying those affordances. Thus, perception and action are coupled cyclically until the intended goal is realized. When affordance perception capacity is disturbed, an observer may be unable to appreciate the rich meanings and values the surrounding environment offers. An affordance perception deficit can trigger

[1] But see Pellicano et al. (2017) for attempts to elucidate apraxia deficits in terms of affordance mechanisms. See also Sevos et al. (2013), who demonstrated that patients with schizophrenia performed poorly in an affordance detection task; and Rounds and Humphreys et al. (2000), who attempted to rationalize limb apraxia arising from an abnormal sensitivity to competition in the presence of multiple affordances. We must note that our understanding of affordance differs from that of Rounds and Humphreys.

a cascade of reactions that lead ultimately to the mental suffering seen in patients with schizophrenia and AD. Gibson (1982) admonished us that "the notion of affordances implies a new theory of meaning" (p. 409). Our findings support a strong argument for exploring the validity of affordance, in particular, and psychological reality, in general, from the perspective of values and meanings.

In conclusion, we suggest that contra-positive evidence of affordance can complement direct evidence for the concept. We suggest further that this integration can enable us to establish a stronger methodological foundation to design research that can help validate affordance empirically and further elucidate the psychological meaning embodied in affordance. In addition, based on our experience, we also suggest that clinical populations, particularly those arising from disturbance of minimal (i.e., ecological) self, can serve as fertile ground from which we can harvest contra-positive evidence, further corroborating Gibson's concept of affordance.

AUTHOR CONTRIBUTIONS

Both authors contributed to manuscript from conception, drafting, revision, read, and approved the submitted version.

FUNDING

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A01061844).

ACKNOWLEDGMENTS

The authors wish to thank Claudia Carello for careful reading of an earlier draft of this manuscript.

REFERENCES

Arango, C., and Carpenter, W. T. (2011). "The schizophrenia construct: symptomatic presentation," in *Schizophrenia*, 3rd Edn, eds D. R. Weinberger and P. J. Harrison (Oxford: Wiley-Blackwell), 9–23.

Ataria, Y., and Horovitz, O. (2021). The destructive nature of severe and ongoing trauma: Impairments in the minimal self. *Philos. Psychol.* 34, 254–276. doi: 10.1080/09515089.2020.1854709

Bajic, V., Misic, N., Stankovic, I., Zaric, B., and Perry, G. (2021). Alzheimer's and consciousness: how much subjectivity is objective? *Neurosci. Insights* 16:26331055211033869. doi: 10.1177/26331055211033869

Baumard, J., Osiurak, F., Lesourd, M., and Le Gall, D. (2014). Tool use disorders after left brain damage. *Front. Psychol.* 5:473. doi: 10.3389/fpsyg.2014.00473

Bomilcar, I., Bertrand, E., Morris, R. G., and Mograbi, D. C. (2021). The seven selves of dementia. *Front. Psychiatry* 12:646050. doi: 10.3389/fpsyt.2021.646050

Cahill, C., and Frith, C. D. (1996). "A cognitive basis for the signs and symptoms of schizophrenia," in *Schizophrenia: A Neuropsychological Perspective,* eds C. Pantelis, H. E. Nelson, and T. R. E. Barnes (New York, NY: John Wiley) 373–395.

Cermolacce, M., Naudin, J., and Parnas, J. (2007). The "minimal self" in psychopathology: re-examining the self disorders in the schizophrenia spectrum. *Conscious. Cogn.* 16, 703–714. doi: 10.1016/j.concog.2007.05.013

de Haan, S. (2020). *Enactive Psychiatry*. Cambridge: Cambridge University Press.

de Haan, S., Rietveld, E., Stokhof, M., and Denys, D. (2013). The phenomenology of deep brain stimulation-induced changes in OCD: an enactive affordance-based model. *Front. Hum. Neurosci.* 7:653. doi: 10.3389/fnhum.2013.00653

Fajen, B. R., and Kim, N.-G. (2002). Perceiving curvilinear heading in the presence of moving objects. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 1100–1119. doi: 10.1037//0096-1523.28.5.1100

Flament-Fultot, M., Nie, L., and Carello, C. (2016). Perception-action mutuality obviates mental construction. *Constr. Found.* 11, 298–307.

Fuchs, T. (2009). Embodied cognitive neuroscience and its consequences for psychiatry. *Poiesis Praxis* 6, 219–233.

Fuchs, T. (2010). The psychopathology of hyperreflexivity. *J. Specul. Philos.* 24, 239–255.

Fuchs, T. (2013). "The self in schizophrenia: jaspers, schneider, and beyond," in *One Century of Karl Jaspers' General Psychopathology*, eds G. Stanghellini and T. Fuchs (Oxford: Oxford University Press), 245–257.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/s1364-6613(99) 01417-5

Gibson, J. J. (1977). "The theory of affordances," in *Perceiving, Acting, And Knowing: Toward an Ecological Psychology*, eds R. E. Shaw and J. Bransford (Hillsdale, NJ: Lawrence Erlbaum Associates), 67–82.

Gibson, J. J. (1979). *The Ecological Approach To Visual Perception*. Boston, MA: Houghton-Mifflin.

Gibson, J. J. (1982). "Notes on affordances," in *Reasons For Realism*, eds E. Reed and R. Jones (Hillsdale, NJ: Lawrence Erlbaum Associates), 401–418.

Heft, H. (2003). Affordances, dynamic experience, and challenge of reification. *Ecol. Psychol.* 15, 149–180.

Heft, H. (2020). Ecological psychology and enaction theory: divergent groundings. *Front. Psychol.* 11:991. doi: 10.3389/fpsyg.2020.00991

Humphreys, G. W., Fode, E. M. E., and Francis, D. (2000). "The organization of sequential actions," in *Attention and Performance XVIII: Control of Cognitive Processes*, eds S. Monsell and J. Driver (Cambridge, MA: MIT Press), 427–442.

Insel, T. R. (2010). Rethinking schizophrenia. *Nature* 468, 187–193.

Jablensky, A. (2010). The diagnostic concept of schizophrenia: its history, evolution, and future prospects. *Dialogues Clin. Neurosci.* 12, 271–287. doi: 10.31887/DCNS.2010.12.3/ajablensky

Kim, N.-G., and Kim, H. (2017). Schizophrenia: an impairment in the capacity to perceive affordances. *Front. Psychol.* 8:1052. doi: 10.3389/fpsyg.2017.01052

Kim, N.-G., Effken, J. A., and Lee, H.-W. (2022). Impaired affordance perception as the basis of tool use deficiency in Alzheimer's disease. *Healthcare* 10:839.

Kyselo, M. (2016). The enactive approach and disorders of the self – the case of schizophrenia. *Phenomenol. Cogn. Sci.* 15, 591–616. doi: 10.1159/000369888

Lee, D. N., and Aronson, E. (1974). Visual proprioceptive control of standing in human infants. *Percept. Psychophys.* 15, 529–532. doi: 10.3758/BF03199297

Lee, D. N., and Lishman, J. R. (1975). Visual proprioceptive control of stance. *J. Hum. Mov. Stud.* 1, 87–95.

Lesourd, M., Baumard, J., Jarry, C., Etcharry-Bouyx, F., Belliard, S., Moreaud, O., et al. (2016). Mechanical problem-solving strategies in Alzheimer's disease and semantic dementia. *Neuropsychology* 30, 612–623. doi: 10.1037/neu0000241

Mark, L. S. (1987). Eyeheight-scaled information about affordances: a study of sitting and stair climbing. *J. Exp. Psychol. Hum. Percept. Perform.* 13, 360–370. doi: 10.1037//0096-1523.13.3.361

Neisser, U. (1988). Five kinds of self-knowledge. *Philos. Psychol.* 1, 35–59.

Neisser, U. (1993). "The self perceived," in *The Perceived Self*, ed. U. Neisser (Cambridge: Cambridge University Press), 3–21.

Nelson, B., and Sass, L. A. (2017). Towards integrating phenomenology and neurocognition: possible neurocognitive correlates of basic self-disturbance in schizophrenia. *Curr. Probl. Psychiatry* 18, 184–200.

Nelson, B., Parnas, J., and Sass, L. A. (2014). Disturbance of minimal self (Ipseity) in schizophrenia: clarification and current status. *Schizophr. Bull.* 40, 479–482. doi: 10.1093/schbul/sbu034

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Parnas, J. (2000). "The self and intentionality in the pre-psychotic stages of schizophrenia," in *Exploring the Self: Philosophical and Psychopathological*

Perspectives on Self-Experience, ed. D. Zahavi (Amsterdam: John Benjamins), 115–147. doi: 10.1075/aicr.23.10par

Parnas, J. (2003). "Self and schizophrenia: a phenomenological perspective," in *The Self In Neuroscience And Psychiatry*, eds T. Kircher and A. David (Cambridge: Cambridge University Press), 217–241.

Parnas, J. (2012). The core gestalt of schizophrenia. *World Psychiatry* 11, 67–69.

Parnas, J., and Sass, L. A. (2010). "Phenomenology of self-disorders," in *The Embodied Self: Dimensions, Coherence And Disorders*, eds T. Fuchs, H. C. Sattel, and P. Henningsen (Stuttgart: Schattauer), 227–244.

Pazzaglia, M., and Galli, G. (2014). Loss of agency in apraxia. *Front. Hum. Neurosci.* 8:751. doi: 10.3389/fnhum.2014.00751

Pellicano, A., Borghi, A. M., and Binkofski, F. (2017). Editorial: bridging the theories of affordances and limb apraxia. *Front. Hum. Neurosci.* 11:148. doi: 10.3389/fnhum.2017.00148

Persons, J. (1986). The advantages of studying psychological phenomena rather than psychiatric diagnoses. *Am. Psychol.* 41, 1252–1260. doi: 10.1037/0003-066X.41.11.1252

Raballo, A., Saebye, D., and Parnas, J. (2011). Looking at the schizophrenia spectrum through the prism of self-disorders: an empirical study. *Schizophr. Bull.* 37, 344–351. doi: 10.1093/schbul/sbp056

Read, C., and Szokolszky, A. (2020). Ecological psychology and enactivism: perceptually-guided action vs. sensation-based enaction. *Front. Psychol.* 11:1270. doi: 10.3389/fpsyg.2020.01270

Reed, E. S. (1996). *Encountering The World*. New York, NY: Oxford University Press.

Sass, L., Borda, J. P., Madeira, L., Pienkos, E., and Nelson, B. (2018). Varieties of self disorder: a bio-pheno-social model of Schizophrenia. *Schizophr. Bull.* 44, 720–727. doi: 10.1093/schbul/sby001

Sass, L. A. (2003a). Negative symptoms, schizophrenia, and the self. *Int. J. Psychol. Ther.* 3, 153–180.

Sass, L. A. (2003b). "Self-disturbance in schizophrenia: hyperreflexivity and diminished self-affection," in *The Self In Neuroscience and Psychiatry*, eds T. Kircher and A. David (Cambridge: Cambridge University Press), 242–271. doi: 10.1159/000488462

Sass, L. A. (2014). Self-disturbance and schizophrenia: structure, specificity, pathogenesis (Current issues, New directions). *Schizophr. Res.* 152, 5–11. doi: 10.1016/j.schres.2013.05.017

Sass, L. A., and Parnas, J. (2003). Schizophrenia, consciousness, and the self. *Schizophr. Bull.* 29, 427–444.

Sevos, J., Grosselin, A., Pellet, J., Massoubre, C., and Brouillet, D. (2013). Grasping the world: object-affordance effect in schizophrenia. *Schizophr. Res. Treat.* 2013:531938. doi: 10.1155/2013/531938

Sherrington, C. S. (1906). *The Integrative Action Of The Nervous System*. New Haven, CT: Yale University Press.

Stanghellini, G. (2009). Embodiment and schizophrenia. *World Psychiatry* 8, 56–59.

Varela, F., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: Oxford University Press.

von Fieandt, K., and Gibson, J. J. (1959). The sensitivity of the eye to two kinds of continuous transformation of a shadow-pattern. *J. Exp. Psychol.* 57, 344–347. doi: 10.1037/h0046028

Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. *J. Exp. Psychol. Hum. Percept. Perform.* 10, 683–703. doi: 10.1037//0096-1523.10.5.683

Warren, W. H. (1998). Visually controlled locomotion: 40 years later. *Ecol. Psychol.* 10, 177–219.

Warren, W. H. (2006). The dynamics of perception and action. *Psychol. Rev.* 113, 358–389.

Warren, W. H. (2021). Information is where you find it: perception as an ecologically well-posed problem. *I Perception* 12, 1–24. doi: 10.1177/20416695211000366

Warren, W. H., and Whang, S. (1987). Visual guidance of walking through apertures: body scaled information for affordances. *J. Exp. Psychol. Hum. Percept. Perform.* 13, 371–383. doi: 10.1037//0096-1523.13.3.371

Weiler, M., Northoff, G., Damasceno, B. P., and Balthazar, M. L. F. (2016). Self, cortical midline structures and the resting state: implications for Alzheimer's disease. *Neurosci. Biobehav. Rev.* 68, 245–255. doi: 10.1016/j.neubiorev.2016.05.028

Weinberger, D. R., and Harrison, P. J. (2011). *Schizophrenia*, 3rd Edn. West Sussex: John Wiley & Sons.

Wong, A. H. C., and Van Tol, H. H. M. (2003). Schizophrenia: from phenomenology to neurobiology. *Neurosci. Biobehav. Rev.* 27, 269–306.

Wraga, M. (1999). The role of eye height in perceiving affordances and object dimensions. *Percept. Psychophys.* 61, 490–507. doi: 10.3758/bf03211968

Zahavi, D. (2005). *Subjectivity And Selfhood: Investigating The First-Person Perspective*. Cambridge, MA: MIT Press.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Kim and Effken. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

SYSTEMATIC REVIEW
published: 29 June 2022
doi: 10.3389/fpsyg.2022.897144

Should Assessments of Decision-Making Capacity Be Risk-Sensitive? A Systematic Review

Noah Clark Berens and Scott Y. H. Kim*

Department of Bioethics, Clinical Center, National Institutes of Health, Bethesda, MD, United States

Background: The concept of decision-making capacity (DMC) or competence remains controversial, despite widespread use. Risk-sensitive DMC assessment (RS-DMC)—the idea that the higher the risk involved in a decision, the greater the decisional abilities required for DMC—has been particularly controversial. We conducted a systematic, descriptive review of the arguments for and against RS-DMC to clarify the debate.

Methods: We searched PubMed/MEDLINE (National Library of Medicine), PsycInfo (American Psychological Association) and Philpapers, updating our search to February 15th, 2022. We targeted peer-reviewed publications in English that argue for or against RS-DMC. Two reviewers independently screened the publications and extracted data from each eligible manuscript.

Results: Of 41 eligible publications, 22 supported a risk-sensitive threshold in DMC assessment. Most arguments for RS-DMC rely on its intuitive appeal and practical merits. The arguments against RS-DMC primarily express concerns about paternalism and the seeming asymmetry between consent and refusal; critics of RS-DMC support epistemic, rather than substantive (i.e., variable threshold), risk-sensitivity; counterarguments responding to criticisms of RS-DMC address charges of paternalism and exhibit a notable variety of responses to the issue of asymmetry. Authors used a variety of frameworks regarding the definition of DMC, its elements, and its relation to decisional authority, and these frameworks were significantly associated with positions on RS-DMC. A limitation of our review is that the coding relies on judgment and interpretation.

Conclusion: The review suggests that some of the debate about RS-DMC stems from differences in underlying frameworks. Most defenses of RS-DMC rely on its intuitive appeal, while most criticisms reflect concerns about paternalism or the asymmetry between consent and refusal. Defenses of RS-DMC respond to the asymmetry problem in a variety of ways. Further research is needed on the implications of underlying frameworks, the asymmetry problem, and the distinction between epistemic and substantive models of RS-DMC.

Keywords: bioethics, decision-making capacity for treatment, review – systematic, capacity, mental competency

OPEN ACCESS

Edited by:
Manuel Trachsel,
University of Zurich, Switzerland

Reviewed by:
Jessica Neri,
University of Padua, Italy
Peter Lepping,
Bangor University, United Kingdom

****Correspondence:***
Scott Y. H. Kim
scott.kim@nih.gov

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *15 March 2022*
Accepted: *03 June 2022*
Published: *29 June 2022*

Citation:
Berens NC and Kim SYH (2022) Should Assessments of Decision-Making Capacity Be Risk-Sensitive? A Systematic Review. Front. Psychol. 13:897144. doi: 10.3389/fpsyg.2022.897144

INTRODUCTION

In most jurisdictions, decision-making capacity (DMC) is used to classify patients into two groups: those whose medical decisions should be made by the patient herself and those whose decisions need to be made by another party. Thus, faulty assessment of DMC can result in either failure to protect a vulnerable incapacitated patient from harm or violation of a capacitated patient's autonomy. Despite its importance, the concept of DMC remains controversial. Issues such as how emotions affect DMC (Charland, 1998), whether the ability to value is relevant to DMC (Kim, 2010), how authenticity should play a role (den Hartogh, 2016; Ahlin Marceta, 2020), and the role of voluntariness have all inspired debate (Charland, 2002). A recent narrative review explores the role of emotions and values, and highlights the complexity and lack of consensus (Hermann et al., 2016).

One particularly controversial point of debate is whether DMC assessment should be risk-sensitive. Risk is a broad term that refers to the seriousness or momentousness of a decision. Thus, risk-sensitive assessment of DMC (RS-DMC) refers to the idea that when the stakes of a decision are high, the level of decisional abilities needed for DMC should be higher as well (Drane, 1984; Buchanan and Brock, 1986; Culver and Gert, 1990). As described in an English legal decision, what matters is whether "[the patient] had a capacity which was commensurate with the gravity of the decision which he purported to make. The more serious the decision, the greater the capacity required" [Re T (Adult: refusal of medical treatment), 1992]. For example, a patient deciding whether or not to withdraw life-sustaining treatment may be held to a high threshold, while a patient deciding between two similarly effective antibiotics may be held to a lower one.

Debate over whether RS-DMC is appropriate seems to have begun in the literature in 1984 (Drane, 1984), yet remains unresolved. Despite persistent disagreement in the literature, RS-DMC is widely accepted and used by clinicians (Kim et al., 2006). The concept has been frequently referenced in UK legal decisions (Buchanan, 2004; Parker, 2006). If RS-DMC is ethically problematic, therefore, it has broad implications for both clinical practice and the law. Furthermore, a wide variety of definitions, concepts, and arguments are used in the literature, making the debate particularly complex. This lack of clarity could lead to inconsistency in DMC evaluation in clinical practice. Accordingly, we conducted a systematic review of the arguments for and against RS-DMC. Our aim is descriptive, so we do not aim to resolve the disagreements about RS-DMC, but to clarify the wide variety of issues at hand in order to promote future fruitful debate.

METHODS

Search Strategy

We used the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) checklist (Page et al., 2021). One author (NB) searched the following citation and abstract databases from inception until August 5th, 2021: PubMed/MEDLINE (US National Library of Medicine), Philpapers, and PsycInfo (American Psychological Association). The search strategy used keywords and controlled vocabulary terms (MeSH and Thesaurus of Psychological index Terms) for the topic of interest. The searches were limited to English language and peer-reviewed publications only where possible. The full search strategy (**Supplementary Materials**) was reviewed and validated by an independent librarian from the National Institutes of Health Library, and the search was updated to February 15th, 2022. Finally, we used the snowball method and our own experience to add publications that were not detected in the three databases. All search results were exported to EndNote X9 and duplicate citations were identified and removed.

Inclusion and Exclusion Criteria

We included an article if:

1) It argues for or against risk-sensitive assessment of decision-making capacity[1]
2) The publication is peer-reviewed
3) The publication is written in English

We excluded purely descriptive publications that do not make an argument for either position. We did not require that publications focus exclusively or predominantly on the issue of RS-DMC, only that they make an argument for or against RS-DMC. Additionally, because some of the initial discussion about RS-DMC originated in books, we eventually included five books that appear in the debates or were known to us in the literature.

Study Selection and Data Extraction

To minimize bias, two readers (NB and a research assistant) independently performed the title/abstract screening and the full-text screening following the predefined inclusion and exclusion criteria. Discrepancies were resolved through discussion until agreement was reached, and when needed, a third reader (SK) assisted in this process. We identified and extracted frameworks and reasons as follows. Both authors (NB, SK) read 50% of the eligible manuscripts and independently created coding schemas. After comparison and discussion, a preliminary schema was agreed upon and adjusted as necessary while coding the eligible manuscripts. The resulting coding scheme had two parts: frameworks, which refer to structural or content features of a publication's conception of concepts related to RS-DMC, and specific reasons used in a publication. Both authors identified passages in each publication that constituted a "reason" or "framework." Reasons were categorized into three groups: (1) Arguments for RS-DMC, (2) Arguments against RS-DMC, and (3) Counterarguments defending RS-DMC[2]. This categorization naturally emerged from the progression of the debate in the literature, as the first publications discussing RS-DMC defended it, and were then critiqued, opening the door for counterarguments. Thus, our categories reflect the progression of

[1] Although the most common context for DMC evaluation is medical, we did not limit our search to this context.

[2] Note that a publication that clearly argues for RS-DMC could still include individual arguments against RS-DMC. Thus, individual arguments attributed to a publication do not necessarily reflect a publication's overall stance on RS-DMC.

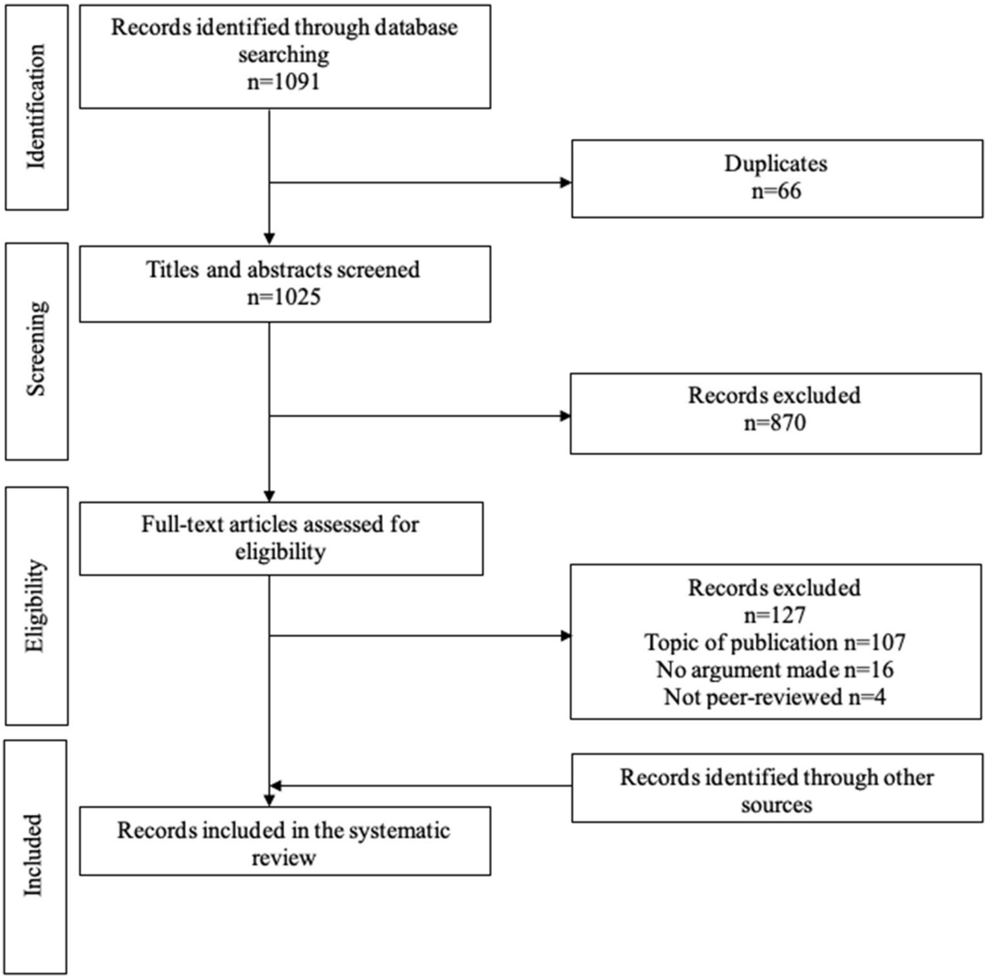

Identification
Records identified through database searching n=1091
Duplicates n=66
Screening
Titles and abstracts screened n=1025
Records excluded n=870
Eligibility
Full-text articles assessed for eligibility
Records excluded n=127
Topic of publication n=107
No argument made n=16
Not peer-reviewed n=4
Records identified through other sources
Included
Records included in the systematic review

FIGURE 1 | PRISMA flow chart for article selection.

the dialogue characteristic of the debate. In addition, we tracked the year of publication, type of journal, and author background.

RESULTS

The systematic search yielded 1,058 articles (**Figure 1**). Of those, 28 were eligible for inclusion. We identified 13 additional eligible publications, for a total of 41 publications (36 articles and 5 books) (Drane, 1984, 1985; Buchanan and Brock, 1986, 1989; Feinberg, 1986; Eastman and Hope, 1988; Kloezen et al., 1988; Culver and Gert, 1990; Brock, 1991; Elliott, 1991; Saks, 1991, 1999; Skene, 1991; Wicclair, 1991a,b, 1999; Winick, 1991; Schopp, 1994; White, 1994; Wilks, 1997, 1999; Grisso and Appelbaum, 1998; Cale, 1999; Maclean, 2000; Berghmans, 2001; Buller, 2001; Checkland, 2001; DeMarco, 2002; Buchanan, 2004; Parker, 2004, 2006; Saks and Jeste, 2006; Howe, 2010; Kim, 2010; Bolt and van Summeren, 2014; Brudney and Siegler, 2015; Manson, 2015; den Hartogh, 2016; Lawlor, 2016; Roberts, 2018; Graber, 2021).

Publication Characteristics

Dates of publication ranged from 1984 and 2021, though most (32/41) were published between 1984 and 2006. Most articles were published in Bioethics/Philosophy journals (20/36), followed by Clinical journals (6/36) and Law/Policy journals (5/36). Nonclinical (philosophy, bioethics, law, policy), was the most common author background (25/41), followed by mixed background (10/41) and clinical (medicine, psychiatry, etc.) background (6/41).

Views on RS-DMC

Three main stances emerged: (1) the substantive view, according to which the threshold required for competence itself should vary with risk, (2) the epistemic view, according to which the amount of evidence or certainty required for a finding of competence should vary with risk, and (3) neither (**Table 1**). A slight majority (22/41) supported the substantive view of RS-DMC, while fewer supported the epistemic view only (12/41). Most publications in clinical journals (5/6) or books (4/5) supported the substantive view. Notably, most articles or books published before 1990 held the substantive view (6/8 publications). All epistemic view articles were published in 1990 or later.

Some publications endorsed both the substantive and epistemic views. These publications were categorized under substantive view, as most of the debate about RS-DMC focuses

TABLE 1 | Views on RS-DMC.

View	Definition	*N*	References
Views on RS-DMC			
View on RS-DMC	*Substantive* The threshold required for competence itself should vary with risk.	22	Drane, 1984, 1985; Buchanan and Brock, 1986, 1989; Feinberg, 1986; Eastman and Hope, 1988; Brock, 1991; Skene, 1991; Winick, 1991; Schopp, 1994; Wilks, 1997, 1999; Grisso and Appelbaum, 1998; Saks, 1999; Berghmans, 2001; Buchanan, 2004; Howe, 2010; Kim, 2010; Bolt and van Summeren, 2014; den Hartogh, 2016; Roberts, 2018
	Epistemic The amount of evidence or confidence required for a finding of competence should vary with risk	12	Wicclair, 1991a,b, 1999; Cale, 1999; Checkland, 2001; DeMarco, 2002; Parker, 2004, 2006; Brudney and Siegler, 2015; Manson, 2015; Lawlor, 2016; Graber, 2021
	Neither Substantive nor Epistemic Does not endorse either view	7	Kloezen et al., 1988; Culver and Gert, 1990; Elliott, 1991; Saks, 1991; White, 1994; Maclean, 2000; Buller, 2001

on whether the substantive view of RS-DMC is appropriate. For instance, articles critical of RS-DMC always targeted the substantive view, never the epistemic. In many cases, articles critical of substantive RS-DMC endorsed the epistemic view. More specifically, the epistemic view is often used to allow for the inclusion of risk in DMC assessment, thus satisfying the common intuition that this is appropriate, while avoiding the supposed flaws of the substantive view. Thus, the debate largely is between those who support the substantive view and those who do not, regardless of whether they support the epistemic view. Accordingly, hereafter we will use 'RS-DMC' to refer to the substantive view of RS-DMC, unless otherwise specified.

Further, there is debate in the literature about the distinction between substantive and epistemic RS-DMC. Some authors question whether these views differ in practice (Wilks, 1999), while others maintain that there is a significant distinction (Wicclair, 1999; Parker, 2004).

Reasons Used in the Debate

We identified 26 primary reasons in the literature that were used in more than one publication (**Tables 2–4**). We classified 8 reasons as "Pros" (**Table 2**), 10 reasons as "Cons" (**Table 3**), and 8 reasons as "Counterarguments Defending RS-DMC" (**Table 4**). This reflects the progression of the debate over time.

Reasons in Favor of RS-DMC

The most common reasons in favor of RS-DMC rely on the intuitive appeal of RS-DMC and its practical merits (**Table 2**). For example, many publications argue that RS-DMC allows for a balance between autonomy and welfare (P2) and the balancing of potential errors (P3), but generally do not explain this balancing in detail, although there are exceptions (Schopp, 1994; Buchanan, 2004). Similarly, "Coheres with Current Practice" (P1), "Rejecting fixed level of competence safeguards against broader paternalism" (P4), "Avoids unnecessarily burdening the system" (P5), and "The opposing view needs to articulate a natural 'adequate level' of decision-making abilities" (P7) focus on the practical difficulties that would result from rejecting RS-DMC. Finally, though cited infrequently, "Respecting patient's wishes has value" (P6) and "Tailored DMC assessment" (P8) do not fall neatly into either intuitive appeal or practical merits.

Reasons Against RS-DMC

Most of the reasons against RS-DMC reflect concerns about paternalism (**Table 3**, arguments C2 through C9). Some arguments are direct charges of paternalism ("RS-DMC is paternalistic" [C2]) while others imply this concern. For example, "Form of outcome-based DMC" (C3) is the argument that when risk is incorporated into DMC assessment, a patient may be deemed incompetent and have their decisions ignored based on merely the decision they make. Similarly, "Imports assessor's values" (C6), "Introduction of values into value-neutral assessment" (C8), and "Falsely finds incompetent persons competent" (C9) are rooted in concerns that DMC assessment will be driven by physicians' values rather than any objective measure, and patients will be judged on whether their decisions/values differ from those of the evaluating physician. Finally, "Conflation of DMC and Decisional Authority" (C4) and "RS-DMC is tautological" (C5) argue that RS-DMC is conceptually flawed and only nominally protects patient autonomy, as a physician could seemingly deem a patient "incompetent" as long as he deemed the risk high enough, rather than assess competence with no consideration of risk and independently decide whether the patient's decision should be respected.

"Standards vary with complexity, not risk" (C7), "Coherence is not sufficient reason" (C10), and "It is unclear where to set the threshold" (C11) do not fit as neatly into this umbrella of concern over paternalism. C10 denies that coherence with current practice is sufficient reason for RS-DMC. C11 highlights a practical concern about RS-DMC implementation. C7 is a disagreement over what explains the intuitive appeal of raising the threshold when decisions are high-risk, but it is disputed by even some critics of RS-DMC.

The asymmetry between consent and refusal (C1) is the most discussed argument regarding RS-DMC. It says that RS-DMC implies the conceptually incoherent view that a person can be competent to consent but not to refuse. Some argue that even if this is not conceptually incoherent, it leads to ethically unacceptable evaluations involving bad faith or deception (Lawlor, 2016).

TABLE 2 | Pro arguments.

Code	Argument type	Argument content and examples	*N*	References
P1	Coheres with current practice	RS-DMC coheres with current medical and/or legal practice and norms or common understanding in every day sense of competence "a concept that allows a raising or lowering of the standard for decision-making capacities depending upon the risks of the decision in question is clearly more consonant with the way people actually make informal competency determinations" (Buchanan and Brock, 1989)	17	Drane, 1985; Buchanan and Brock, 1986, 1989; Feinberg, 1986; Brock, 1991; Skene, 1991; Winick, 1991; Schopp, 1994; Wilks, 1997, 1999; Grisso and Appelbaum, 1998; Buchanan, 2004; Howe, 2010; Kim, 2010; Lawlor, 2016; Graber, 2021
P2	Balances autonomy and welfare	RS-DMC is the best way to balance the competing values of autonomy/self-determination and well-being/welfare "It allows a better balance between the competing values of self-determination and well-being that are to be served by a determination of competence" (Buchanan and Brock, 1986)	14	Drane, 1984, 1985; Buchanan and Brock, 1986, 1989; Eastman and Hope, 1988; Brock, 1991; Winick, 1991; Grisso and Appelbaum, 1998; Berghmans, 2001; Kim, 2010; Bolt and van Summeren, 2014; Brudney and Siegler, 2015; den Hartogh, 2016; Lawlor, 2016
P3	Balances potential errors	RS-DMC balances two potential errors: (1) authorizing an incompetent patient's decision, leading to harm and (2) overruling a competent patient's decision, disrespecting their autonomy. As risk increases, (1) is more damaging than (2), requiring that the standard for deeming a patient competent increases. "A properly performed competency assessment should eliminate two types of error: (1) preventing a competent person from participating in treatment decisions and (2) failing to protect an incompetent person from the harmful effects of a bad decision." (Drane, 1984)	7	Drane, 1984, 1985; Buchanan and Brock, 1986, 1989; Schopp, 1994; Berghmans, 2001; Buchanan, 2004
P4	Rejecting fixed level of competence safeguards against broader paternalism	If there is one fixed level of competence that applies to all situations, it has broader paternalistic consequences. "the alternative would be to let go of the presumption of competence itself, and examine patients' competence in all cases, whether or not there is any reason for doubt. That would really be paternalistic in the extreme" (den Hartogh, 2016).	5	Drane, 1984; Buchanan and Brock, 1986, 1989; Winick, 1991; den Hartogh, 2016
P5	Avoids unnecessarily burdening the system	If every impaired person is interrogated or held to some high standard relative to the risk, then system would be burdened with very little gained. "it … would expend scarce resources without achieving significant benefits" (Winick, 1991)	4	Drane, 1984, 1985; Winick, 1991; Brudney and Siegler, 2015
P6	Respecting patients' wishes has value	RS-DMC allows for lowering the standard for DMC and respecting patient's wishes in low-risk situations, which is valuable "applying a lower standard of competency when a patient assents to a recommended course of treatment than when a patient objects, serves not only individual autonomy values, but also the interest in promoting health." (Winick, 1991)	2	Winick, 1991; Buchanan, 2004
P7	The opposing view needs to articulate a natural 'adequate level' of decision-making abilities	If risk is not used in setting a threshold for DMC, a fixed standard must be identified and defended, but no such model exists. "On both views a certain point on the scale of competence can be identified at which we are prepared to attribute that authority in a particular case, even if, on the multi-dimensional view, we also have to take other considerations into account in order to do that. How do we identify that point? I will argue that until now only the multi-dimensional theory has been able to provide a plausible answer to that question." (den Hartogh, 2016)	2	Brock, 1991; den Hartogh, 2016
P8	Tailored DMC assessment	RS-DMC allows for DMC assessment to be tailored to the needs of each patient "The use of a sliding scale allows care providers to tailor the standard they use to the particular needs of each patient" (Howe, 2010)	2	Howe, 2010; Bolt and van Summeren, 2014

TABLE 3 | Con arguments.

Code	Argument type	Argument content and examples	*N*	Authors represented
C1	Asymmetry between consent and refusal	Asymmetry between consent and refusal is conceptually incoherent or problematic "Extant accounts of risk- related standards of capacity appear to be committed to the existence of asymmetrical capacity, that is, cases where a patient is capacitated to accept treatment but lacks capacity to reject treatment. However, asymmetrical capacity appears to be conceptually incoherent; in such cases, there is no sense to be made of the claim that the patient either has, or lacks, capacity." (Graber, 2021)	11	Culver and Gert, 1990; Wicclair, 1991a,b, 1999; Cale, 1999; Maclean, 2000; Berghmans, 2001; Buller, 2001; Manson, 2015; Lawlor, 2016; Graber, 2021
C2	RS-DMC is paternalistic	RS-DMC is inherently paternalistic and inconsistent with autonomy, or is highly prone to paternalistic abuse by allowing evaluator to set threshold according to their own values "[T]here is a danger that standards of understanding, reasoning, and so forth will be set arbitrarily and unattainably high by those who believe that paternalism is justified when perceived risks are great." (Wicclair, 1991a)	11	Culver and Gert, 1990; Saks, 1991; Wicclair, 1991a,b; White, 1994; Cale, 1999; Maclean, 2000; Berghmans, 2001; DeMarco, 2002; Buchanan, 2004; Parker, 2004
C3	Form of outcome-based DMC	DMC assessment should be process-oriented and should not depend on the likely outcome of the choice an individual makes. "their account appears to be incompatible with the principle that assessments of decision-making capacity should utilize a standard that is process-oriented, and not result-oriented." (Wicclair, 1991b)	9	Wicclair, 1991a,b; White, 1994; Cale, 1999; Saks, 1999; Maclean, 2000; Buller, 2001; Parker, 2004; Saks and Jeste, 2006
C4	Conflation of DMC and DA	RS-DMC conflates two distinct judgments: (1) whether a person has DMC/competence and (2) whether their decision should have authority "The sliding-scale model of competence based on risk conflates two different questions: (1) whether the patient is competent, and (2) whether we should respect the patient's decision" (Elliott, 1991)	8	Culver and Gert, 1990; Elliott, 1991; Wicclair, 1991a, 1999; White, 1994; Berghmans, 2001; Buller, 2001; DeMarco, 2002
C5	Respecting competent patient's decision is a tautology	If a variable standard is used, prohibition of paternalism (overriding a competent patient's decision) is a mere tautology rather than a strong commitment to patient autonomy "Since the statement that the treatment preferences of competent patients are not to be set aside for paternalistic reasons amounts to a tautology, it hardly reflects a strong commitment to the ethical principle that treatment choices of autonomous patients should be respected." (Wicclair, 1991a)	8	Culver and Gert, 1990; Elliott, 1991; Wicclair, 1991a,b, 1999; Maclean, 2000; DeMarco, 2002; den Hartogh, 2016
C6	Imports assessor's values	RS-DMC relies on the assessor's judgment and values over those of the patient "[T]his manner of assessing competency allows the evaluator to determine that a choice is problematic based upon his or her own values" (Saks, 1999)	8	Saks, 1991, 1999; White, 1994; Cale, 1999; Maclean, 2000; Parker, 2004; Saks and Jeste, 2006; Manson, 2015
C7	Standards vary with complexity, not risk	Complexity of decisions, not risk, explains our intuitions about high-risk decision-making "There may be a correlation between greater risk and increased complexity of requisite decision making skills and abilities" (Wicclair, 1999)	5	Kloezen et al., 1988; Wicclair, 1991a, 1999; Maclean, 2000; Berghmans, 2001
C8	Introduction of values into value-neutral assessment	DMC assessment should be value neutral, but RS-DMC introduces normative values into assessment "understanding competence as related to outcomes requires the unjustified imposition of normative values in the assessment of competence, thereby confusing the kind of competence that a standard is aimed at assessing" (Cale, 1999)	4	White, 1994; Cale, 1999; DeMarco, 2002; Parker, 2004
C9	RS-DMC falsely finds incompetent persons competent	RS-DMC allows those who lack the abilities required to make decisions to be deemed competent or accountable in low-risk situations "As a result, there is the danger that decision-making standards will be set so low when patients concur with the recommendations of health care professionals that they will be classified as decisionally capable, regardless of their mental status." (Wicclair, 1991a)	4	Wicclair, 1991a,b; Berghmans, 2001; Lawlor, 2016

(Continued)

TABLE 3 | Continued

Code	Argument type	Argument content and examples	*N*	Authors represented
C10	Coherence is not sufficient reason	Legal or medical coherence is not a good reason, or the status quo is problematic "At any rate, it seems a poor reason to adopt a misleading definition of a concept to say it accords better with a legal tradition that is itself vague and confused." (Culver and Gert, 1990)	3	Culver and Gert, 1990; Saks, 1991; DeMarco, 2002
C11	It is unclear where to set the threshold	When risk is included, it is unclear where the threshold for DMC should be set "First, let us note that the question of what different standards of capacity would actually look like never arises in most of the risk-related accounts. All we hear is that in cases of higher risk, a higher standard of decision-making capacity is required." (Parker, 2004)	2	Kloezen et al., 1988; Parker, 2004

Counterarguments Against Criticisms of RS-DMC

Most of the counterarguments defend RS-DMC against charges of paternalism (**Table 4**). Some do so by directly refuting the criticisms ("RS-DMC is not paternalistic," [CA2], "RS-DMC is not tautological," [CA5]), but others defend RS-DMC by clarifying how the model functions in practice or by disputing the conceptual premises of the critics. For example, "Outcome alone does not determine DMC" (CA3) clarifies that RS-DMC may include risk or outcome as *part* of the assessment, but does not rely on these factors alone; therefore, outcome alone does not determine DMC as is suggested by critics. Similarly, "Consistent with the reasonable person standard" (CA6) points out how consideration of risk is consistent with the commonly used reasonable person standard, and thus does not import values inappropriately. "Not a conflation" (CA4) argues that the disagreement is due to differences in the conceptual frameworks of decision-making capacity used. Similarly, "Inherently normative" (CA8) asserts the position that DMC assessment cannot be value-neutral.

The counterarguments against the asymmetry argument (CA1), however, are quite varied. Some simply acknowledge that the asymmetry seems odd but embrace it as part of RS-DMC (Buchanan and Brock, 1986; Howe, 2010), or consider consenting and refusing two separate decisions[3] (Brock, 1991; Wilks, 1997). Others show that asymmetry appears acceptable when determining whether to maintain or rebut the presumption of capacity in a given situation (Checkland, 2001; Brudney and Siegler, 2015). Other supporters of RS-DMC argue further that aside from the context of evaluating the presumption of capacity, RS-DMC not only does not need but should not include asymmetry of consent and refusal (Bolt and van Summeren, 2014; den Hartogh, 2016; Graber, 2021).

Underlying Frameworks Used by Authors

Authors varied on how competence and related concepts are defined or understood. We use "frameworks" to refer to structural or content features of the author's conception of competence and related concepts (**Table 5**). We categorized conservatively: a publication was a given category only if it explicitly endorsed or clearly made use of a particular framework or definition; thus, it is possible an author in fact holds a certain framework but we could not code a publication as such.

The first framework category of "Externalism" vs. "Internalism" is a distinction noted by some authors (Wilks, 1997; Berghmans, 2001). Internalism holds that competence is solely a function of a person's internal abilities relevant to decision-making; externalism holds that competence is determined by both internal abilities and other contextual or relational factors external to the person's decisional abilities, such as risk. Of those publications codable on this issue, most were externalist (17/26).

A closely related framework category—"One-step" vs. "Two-step"—addresses how decisional authority should be determined (Buchanan and Brock, 1986; Culver and Gert, 1990). The two options are either a single step that incorporates information about a patient's abilities in addition to relevant contextual factors such as risk or taking two steps, first determining competence, then separately determining whether the person should have decisional authority. Of the 19 publications codable on this issue, most held the one-step view (12/19).

Similarly, the third framework category asks, "Does having DMC imply having decisional authority?" Some hold that a finding of DMC grants an individual decisional authority, so their decisions must be respected, while others hold that a finding of DMC only means that an individual has the abilities to make decisions, not that they should automatically have decisional authority. The majority (17/24) endorsed the view that having DMC implies having decisional authority. The first three framework categories seem closely related conceptually.

The fourth framework category is the *predominant* conception of well-being that a publication relies on in its model of DMC. There were three coding options: (1) predominantly objective, when publications primarily use an objective or shared understanding of welfare, (2) predominantly

[3] The question of whether or not consenting and refusing should be considered two separate decisions appears in the law as well. For example, the Mental Capacity Act of 2005 in England and Wales states that "the courts do not examine separately capacity to consent and capacity to refuse medical treatment. Rather, the courts proceed by examining the question of whether the person has the capacity to make a decision in relation to the treatment.

TABLE 4 | Counterarguments.

Code	Argument type	Argument content and examples	*N*	Authors represented
CA1*	Asymmetry is not problematic or needed	(a) Asymmetry is admittedly odd but cost is acceptable; "There is an important implication of this view that the standard of competence ought to vary with the expected harms or benefits to the patient of acting in accordance with a choice–namely, that just because a patient is competent to consent to a treatment, it does not follow that the patient is competent to refuse it, and vice versa." (Buchanan and Brock, 1986) (b) Consenting and refusing are separate decisions, so there is no asymmetry in RS-DMC "One reason a patient might be competent to consent but not to refuse a treatment, and vice versa, is that the two choices to consent or refuse will be based on different processes of reasoning or decision-making; the overall processes of reasoning must be different if for no other reason than that they result in different choices." (Brock, 1991) (c) Asymmetry is about presumption of capacity, and is actually justified "the greater the risk to the patient, the more reason the physician has to think about capacity." (Brudney and Siegler, 2015) (d) If (c) is accepted, then asymmetry is not needed for RS-DMC "It may be that if the patient consents there is no reason to investigate his competence, but if he refuses, there is. However, if the conclusion following from that investigation is negative, it holds for the consent as well as for the refusal." (den Hartogh, 2016)	15	Buchanan and Brock, 1986, 1989; Brock, 1991; Winick, 1991; Wilks, 1997, 1999; Berghmans, 2001; Checkland, 2001; Howe, 2010; Kim, 2010; Bolt and van Summeren, 2014; Brudney and Siegler, 2015; den Hartogh, 2016; Lawlor, 2016; Graber, 2021
CA2	RS-DMC is not paternalistic	Any argument that claims RS-DMC is not paternalistic "it will generally be the case that if the patient's decision does not coincide with the opinion of the physician, this may trigger the need to assess the patient's capacity. This, however, should not be confused with 'lowering the bar' for incapacity." (Berghmans, 2001)	13	Drane, 1984, 1985; Buchanan and Brock, 1986, 1989; Feinberg, 1986; Eastman and Hope, 1988; Winick, 1991; Schopp, 1994; Wilks, 1997; Berghmans, 2001; Kim, 2010; Roberts, 2018; Graber, 2021
CA3	Outcome alone does not determine DMC	RS-DMC may include the outcome/choice itself as indicative of risk, but other factors are also essential to DMC assessment. "But outcome is not the standard of competence in this model. Rather it is an important factor in only one class of medical decisions." (Drane, 1985)	10	Drane, 1985; Buchanan and Brock, 1986, 1989; Eastman and Hope, 1988; Winick, 1991; Schopp, 1994; Wilks, 1997; Saks, 1999; Berghmans, 2001; Buchanan, 2004
CA4	Not a conflation but different framework of DA	RS-DMC is a conflation of DMC and DA only if you believe DMC is purely a matter of abilities; if you accept that the function of DMC assessment is to determine DA, it is not a conflation "Wicclair insists our account conflates two distinct questions - first, is the patient competent to make the decision and, second, is there reason to disregard the patient's decision and have a surrogate decide for the patient. As we discussed (65–70), an alternative account of competence is possible in which these two questions are distinguished. Competence would then be understood as requiring some minimum threshold of decisionmaking capacities, though the threshold could still be decision-specific and variable, but not as determining decisional authority. This account would leave open whether a patient's competent choice should be set aside on paternalistic grounds in order to protect his or her well-being. We called this a two-step model of patient decision-making authority. We evaluated such a model and argued that our own account was preferable" (Brock, 1991)	5	Buchanan and Brock, 1989; Brock, 1991; Skene, 1991; Berghmans, 2001; Bolt and van Summeren, 2014
CA5	RS-DMC is not tautological	Any argument that responds to the criticism that RS-DMC makes respecting a competent patient's decision tautological "Is our view problematic and empty of any commitment to individual self-determination in this way? It would be if we offered no other criteria for a justified finding of incompetence than that others believed setting aside patients' treatment choices for their own good was justified." (Brock, 1991)	2	Brock, 1991; Wilks, 1997

(Continued)

TABLE 4 | Continued

Code	Argument type	Argument content and examples	*N*	Authors represented
CA6	RS-DMC is consistent with the reasonable person standard	Risk consideration does not import assessor's values, as it is consistent with the commonly accepted 'reasonable person standard'. "[T]reatment refusal does reasonably raise the question of a patient's competence in a way that acceptance of recommended treatment does not. It is a reasonable assumption that physicians' treatment recommendations are more often than not in the interests of their patients. Consequently, it is a reasonable presumption-though rebuttable in any particular instance- that a treatment refusal is contrary to the patient's interest." (Buchanan and Brock, 1986)	10	Drane, 1985; Buchanan and Brock, 1986, 1989; Feinberg, 1986; Skene, 1991; Winick, 1991; Brudney and Siegler, 2015; Lawlor, 2016; Graber, 2021
CA7	Complexity alone can't explain variable standard	Riskier decisions are not necessarily more complex, so risk itself must be what is responsible for the intuitive appeal of variable thresholds "However, complexity is not the same thing as risk: a high-risk procedure may be extremely straightforward, and a low-risk procedure could be quite complicated." (Parker, 2004)	7	Brock, 1991; Skene, 1991; White, 1994; Wilks, 1997; Buller, 2001; Parker, 2004; Kim, 2010
CA8#	DMC assessment is inherently normative	It is impossible for DMC assessment to be value-neutral; it naturally relies on normative judgments "Given the uncertainty and inherent vagueness of the criteria to be applied, physicians assigned the task of assessing competency inevitably make normative judgments." (Winick, 1991)	10	Winick, 1991; Wilks, 1997, 1999; Grisso and Appelbaum, 1998; Saks, 1999; Berghmans, 2001; Saks and Jeste, 2006; Kim, 2010; Bolt and van Summeren, 2014; den Hartogh, 2016

*Counterarguments are numbered the same as the con argument to which they respond.

#We did not find any counterarguments directly responding to C9, C10, or C11.

subjective, when publications primarily use an individual's own subjective understanding of their welfare, and (3) both objective and subjective used, with neither clearly favored over the other. Among codable documents on this issue, half (15/30) had a predominantly objective conception, while 9/30 relied on both, and the remaining 6/30 relied primarily on a subjective perspective.

Finally, publications were categorized according to their view on the scope of competence, that is, whether competence is about a *specific, particular* decision or about a *type* of decision. Those publications that hold the first view argue that the specific decision a person makes is relevant to competence assessment, while those that hold the second view argue that an individuals' competence should instead depend on an individual's ability to make the relevant *type* of decision. 17/24 codable publications on this issue supported the "specific decision" view.

Relationship of Frameworks to Stance on RS-DMC

The framework categories identified in **Table 5** are highly associated with an author's stance on RS-DMC.

For example, 16 of 17 publications that hold an externalist view endorse the substantive view of RS-DMC, while all 9 publications that hold an internalist view do not (**Table 6**). Similarly, all 12 publications that support a one-step determination of decisional authority endorse the substantive view, 12 out of 15 publications that use an objective conception of well-being support the substantive view, and 14 out of 17 publications that use "competence of specific decision" support the substantive view.

Relationship of Frameworks to Reasons

The frameworks are sometimes also associated with reasons for or against RS-DMC. This was most obvious when there was a logical connection between the frameworks and the reasons. For example, the two closely inter-related framework elements of "two step vs. one step" view of decisional authority and whether DMC implies having decisional authority were highly associated with two arguments against RS-DMC, namely, whether an author criticized RS-DMC as conflating DMC with decisional authority (C4) and as providing only tautological prohibition of paternalism (C5). For example, 5/7 publications that hold that having DMC does not imply having decisional authority argue that RS-DMC involves a conflation of DMC and decisional authority (C4), whereas authors who view DMC as implying decisional authority understandably do not see a conflation (0 among 12 codable papers). Similarly, 6/7 publications that hold a two-step view of DMC assessment and 6/7 publications that hold that having DMC does not imply having decisional authority argue that RS-DMC prohibits paternalism by definition only ("RS-DMC is tautological" [C5]). These findings suggest that at least some areas of debate over RS-DMC arise due to differing underlying premises.

DISCUSSION

The concept of DMC is widely used every day in most jurisdictions, yet it still engenders debate and disagreement. One particularly controversial debate is whether DMC assessment should be risk-sensitive. This debate began in earnest in 1984, yet remains controversial. Our review of the arguments used in the debate reveal several key findings.

TABLE 5 | Frameworks used.

Framework	Definition	*N*	References
Externalist or Internalist	*Externalist* Competence judgment determined by both internal abilities and other contextual or relational factors external to the abilities.	17	Buchanan and Brock, 1986, 1989; Feinberg, 1986; Eastman and Hope, 1988; Brock, 1991; Winick, 1991; Wilks, 1997, 1999; Grisso and Appelbaum, 1998; Berghmans, 2001; Buchanan, 2004; Saks and Jeste, 2006; Kim, 2010; Bolt and van Summeren, 2014; den Hartogh, 2016; Lawlor, 2016; Roberts, 2018
	Internalist Competence judgment determined solely by level of abilities within the person	9	Kloezen et al., 1988; Culver and Gert, 1990; Wicclair, 1991a,b, 1999; White, 1994; Cale, 1999; Maclean, 2000; Checkland, 2001
One step or two step	*One step* Arriving at DA judgment is a single step that incorporates information about P's abilities plus contextual factors (e.g., risk).	12	Buchanan and Brock, 1986, 1989; Feinberg, 1986; Eastman and Hope, 1988; Brock, 1991; Skene, 1991; Winick, 1991; Grisso and Appelbaum, 1998; Buchanan, 2004; Kim, 2010; Bolt and van Summeren, 2014; den Hartogh, 2016
	Two step Evaluators should assess an individual's abilities in order to reach a competence judgment; then, decide whether the person has decisional authority	7	Culver and Gert, 1990; Elliott, 1991; Wicclair, 1991a,b, 1999; Maclean, 2000; Buller, 2001
Does having DMC imply having DA?	*Yes* A finding of DMC gives an individual DA, so their decisions should be respected	17	Buchanan and Brock, 1986, 1989; Feinberg, 1986; Eastman and Hope, 1988; Kloezen et al., 1988; Brock, 1991; Skene, 1991; Winick, 1991; White, 1994; Wilks, 1997; Grisso and Appelbaum, 1998; Checkland, 2001; Buchanan, 2004; Kim, 2010; Bolt and van Summeren, 2014; den Hartogh, 2016; Roberts, 2018
	No A finding of DMC means the person has the ability to make decisions. Whether their decision should be respected is a separate judgment.	7	Elliott, 1991; Wicclair, 1991a,b, 1999; Maclean, 2000; Buller, 2001
Conception of well-being	*Objective* Author emphasizes or uses objective or shared meaning of welfare.	15	Drane, 1984, 1985; Eastman and Hope, 1988; Culver and Gert, 1990; Winick, 1991; Schopp, 1994; Wilks, 1997; Grisso and Appelbaum, 1998; Buchanan, 2004; Saks and Jeste, 2006; Kim, 2010; Bolt and van Summeren, 2014; Manson, 2015; den Hartogh, 2016; Lawlor, 2016
	Subjective Author emphasizes or uses individual subject's own meaning of welfare or value.	6	Buchanan and Brock, 1986; White, 1994; Grisso and Appelbaum, 1998; Cale, 1999; Buller, 2001; Roberts, 2018; Graber, 2021
	Both	9	Feinberg, 1986; Skene, 1991; Saks, 1999; Wilks, 1999; Maclean, 2000; DeMarco, 2002; Howe, 2010
Specific decision or type of decision	*Specific decision* The specific decision a person makes is relevant to competence assessment	17	Drane, 1984, 1985; Buchanan and Brock, 1986; Feinberg, 1986; Eastman and Hope, 1988; Brock, 1991; Winick, 1991; Grisso and Appelbaum, 1998; Wicclair, 1999; Berghmans, 2001; Saks and Jeste, 2006; Howe, 2010; Kim, 2010; Bolt and van Summeren, 2014; den Hartogh, 2016; Lawlor, 2016; Graber, 2021
	Type of decision Competence assessment is about the person's more general decision-making abilities, not the specific decision they make.	7	Kloezen et al., 1988; Culver and Gert, 1990; Saks, 1991, 1999; Cale, 1999; Maclean, 2000; Buller, 2001

The Importance of Frameworks

There is a lack of uniformity in vocabulary and definitions in the debate. The publications were in journals from a variety of disciplines, by authors with diverse backgrounds and across jurisdictions, which may explain some of the differences in vocabulary. But authors often understand the concepts of DMC and decisional authority differently, as captured by the five framework categories that we tracked in the literature (**Table 5**).

These differences can make it unclear whether disagreement reflects misunderstanding or substantive ethical disagreement. It appears at least some of the disagreements about RS-DMC may be due to differences in frameworks and premises[4]. The frameworks endorsed also sometimes relate to the reasons each publication used. For example, "Conflation of DMC and decisional authority" (C4) or the tautology argument (C5) require the view that having DMC does not imply having decisional authority, a minority view in the literature. The significance of frameworks has been relatively neglected in the debate. Future debates on RS DMC may benefit from explicit attention to this issue.

Patterns in Reasons Used

Though there are many distinct reasons for and against RS-DMC in the literature, a few broad patterns became apparent. First, the

[4] Does one's view on RS-DMC determine one's framework elements or vice versa? This is not an easy question to answer. For example, although it is true that RS-DMC logically implies an externalist view of DMC and therefore one could argue that externalism/internalism element is an implication rather than a premise of RS-DMC, it is not so obvious whether RS-DMC requires a specific view about whether DMC implies decisional authority.

TABLE 6 | Relationship of frameworks to stance on RS-DMC.

Framework	View	Substantive % (*N*)	Epistemic only % (*N*)	Neither S nor E % (*N*)
Externalism vs. Internalism	Externalist view of DMC	94.1 (16/17)	5.9 (1/17)	0 (0/17)
	Internalist view of DMC	0 (0/9)	55.6 (5/9)	44.4 (4/9)
	Uncoded	40 (6/15)	40 (6/15)	20 (3/15)
One-step vs. Two-step	One-step determination of DA	100 (12/12)	0 (0/12)	0 (0/12)
	Two-step determination of DA	0 (0/7)	42.9 (3/7)	57.1 (4/7)
	Uncoded	45.5 (10/22)	40.9 (9/22)	13.6 (3/22)
Does having DMC imply having decisional authority?	Having DMC implies having decisional authority	82.4 (14/17)	5.9 (1/17)	11.8 (2/17)
	Having DMC does not imply having decisional authority	0 (0/7)	42.9 (3/7)	57.1 (4/7)
	Uncoded	47.1 (8/17)	47.1 (8/17)	5.9 (1/17)
Conception of wellbeing	Objective wellbeing	80 (12/15)	13.3 (2/15)	6.7 (1/15)
	Subjective wellbeing	33.3 (2/6)	33.3 (2/6)	33.3 (2/6)
	Both	66.7 (6/9)	22.2 (2/9)	11.1 (1/9)
	Uncoded	18.2 (2/11)	54.5 (6/11)	27.3 (3/11)
Specific decision or type of decision	Competence of specific decision	82.4 (14/17)	17.6 (3/17)	0 (0/17)
	Competence of type of decision	14.3 (1/7)	14.3 (1/7)	71.4 (5/7)
	Uncoded	41.2 (7/17)	47.1 (8/17)	11.8 (2/17)

pro RS-DMC arguments tend to rely on the intuitive and practical appeal of RS-DMC. Second, arguments against RS-DMC mostly have to do with two concerns: one, concerns about paternalism (although sometimes this is only implicit) and, two, concern about the coherence of asymmetry of consent and refusal that is said to be part of RS-DMC. Finally, the most notable feature of the counterarguments defending RS-DMC (aside from defending against the variety of charges of paternalism) is that there were a variety of responses to the asymmetry argument with differing views among defenders of RS-DMC. Given that the RS-DMC debate has perhaps focused more on the asymmetry argument than any other issue, this is an interesting finding and suggests that further research is needed.

Future Work

In addition to the issue of asymmetry, future work should focus more on the distinction between the substantive and epistemic views of RS-DMC. It is curious that those critical of RS-DMC often permit the incorporation of risk into DMC assessment through the epistemic view, rather than arguing risk should be entirely irrelevant. Thus, the intuitive appeal of risk-sensitivity carries a significant weight even among interlocutors who disagree with the substantive view of RS-DMC.

Additionally, it is notable that no publications critical of risk-sensitive DMC assessment targeted the epistemic view. However, whether there is a practical difference between the epistemic vs. substantive view of RS-DMC is disputed (Wicclair, 1999; Wilks, 1999; Parker, 2004). Given how differently each view is treated in the debate, clarification of the precise differences between the views would be valuable.

Limitations

Both construction and application of codes require judgment and interpretation. For example, some codes have significant conceptual overlap—e.g., "Does having DMC imply having decisional authority" and "Two-step vs. One-step"—but we felt it was important to track both codes separately to capture their nuances, particularly because, in order to minimize bias, we coded conservatively and only coded a reason when it was explicit. Our inclusion of books in addition to our systematic search of articles is a further limitation, as a systematic search of books was not possible. However, it seems unlikely that this led to our missing any major arguments or reasons for or against RS-DMC. Only publications written in English were included in our search, and most of the publications included are from the US or UK. Literature in other languages or published in other countries may provide different perspectives on RS-DMC. For example, the emphasis on autonomy may vary among different cultures (Lepping and Raveesh, 2014), and this could also affect views regarding RS-DMC.

CONCLUSION

Whether assessment of DMC should be risk-sensitive is an important and hotly contested issue. We find that some of the debate stems from differences in underlying conceptual frameworks of the authors, as the frameworks are highly associated with one's stance on RS-DMC. Most positive defenses of RS-DMC rely on its intuitive appeal, while most criticisms are driven by concern about paternalism or the asymmetry between consent and refusal. It is notable that defenders of RS-DMC address the asymmetry concern in a variety of ways, suggesting that more attention to this issue is needed. Future work should also clarify the differences between the epistemic and substantive views of RS-DMC.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

NB and SK both contributed to project conception and design, screening and coding of manuscripts, and preparing and editing manuscript drafts. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported in part by the Intramural Research Program of the National Institutes of Health Clinical Center (CL010542).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.897144/full#supplementary-material

REFERENCES

Ahlin Marceta, J. (2020). Resolved and unresolved bioethical authenticity problems. *Monash Bioeth. Rev.* 38, 1–14. doi: 10.1007/s40592-020-00108-y

Berghmans, R. L. P. (2001). Capacity and consent. *Curr. Opin. Psychiatry* 14, 491–499. doi: 10.1097/00001504-200109000-00012

Bolt, I. L., and van Summeren, M. J. (2014). Competence assessment in minors, illustrated by the case of bariatric surgery for morbidly obese children. *Best Pract. Res. Clin. Gastroenterol.* 28, 293–302. doi: 10.1016/j.bpg.2014.02.006

Brock, D. W. (1991). Decisionmaking competence and risk. *Bioethics* 5, 105–112. doi: 10.1111/j.1467-8519.1991.tb00151.x

Brudney, D., and Siegler, M. (2015). A justifiable asymmetry. *J. Clin. Ethics* 26, 100–103. doi: 10.4236/ns.2015.72011

Buchanan, A. (2004). Mental capacity, legal competence and consent to treatment. *J. R. Soc. Med.* 97, 415–420. doi: 10.1177/014107680409700902

Buchanan, A., and Brock, D. W. (1986). Deciding for others. *Milbank Q* 64(Suppl. 2), 17–94. doi: 10.2307/3349960

Buchanan, A. E., and Brock, D. W. (1989). *Deciding for Others: The Ethics of Surrogate Decision Making*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9781139171946

Buller, T. (2001). Competency and risk-relativity. *Bioethics* 15, 93–109. doi: 10.1111/1467-8519.00218

Cale, G. S. (1999). Risk-related standards of competence: continuing the debate over risk-related standards of competence. *Bioethics* 13, 131–148. doi: 10.1111/1467-8519.00137

Charland, L. C. (1998). Appreciation and emotion: theoretical reflections on the MacArthur Treatment Competence Study. *Kennedy Inst. Ethics J.* 8, 359–376. doi: 10.1353/ken.1998.0027

Charland, L. C. (2002). Cynthia's dilemma: consenting to heroin prescription. *Am. J. Bioethics* 2, 37–47. doi: 10.1162/152651602317533686

Checkland, D. (2001). On risk and decisional capacity. *J. Med. Philos.* 26, 35–59. doi: 10.1076/jmep.26.1.35.3035

Culver, C. M., and Gert, B. (1990). The inadequacy of incompetence. *Milbank Q* 68, 619–643. doi: 10.2307/3350196

DeMarco, J. P. (2002). Competence and paternalism. *Bioethics* 16, 231–245. doi: 10.1111/1467-8519.00283

den Hartogh, G. (2016). Do we need a threshold conception of competence? *Med. Health Care Philos.* 19, 71–83. doi: 10.1007/s11019-015-9646-5

Drane, J. F. (1984). Competency to give an informed consent. a model for making clinical assessments. *JAMA* 252, 925–927. doi: 10.1001/jama.1984.03350070043021

Drane, J. F. (1985). The many faces of competency. *Hastings Cent. Rep.* 15, 17–21. doi: 10.2307/3560639

Eastman, N. L., and Hope, R. A. (1988). The ethics of enforced medical treatment: the balance model. *J. Appl. Philos.* 5, 49–59. doi: 10.1111/j.1468-5930.1988.tb00228.x

Elliott, C. (1991). Competence as accountability. *J. Clin. Ethics* 2, 167–71.

Feinberg, J. (1986). *Harm to Self: The Moral Limits of the Criminal Law*. Oxford: Oxford University Press Inc.

Graber, A. (2021). Justifying risk-related standards of capacity via autonomy alone. *J. Med. Ethics.* 47, e89. doi: 10.1136/medethics-2020-106733

Grisso, T., and Appelbaum, P. S. (1998). *Assessing Competence to Consent to Treatment: A Guide for Physicians and Other Health Professionals*. London: Oxford University Press.

Hermann, H., Trachsel, M., Elger, B. S., and Biller-Andorno, N. (2016). Emotion and value in the evaluation of medical decision-making capacity: a narrative review of arguments. *Front. Psychol.* 7, 765. doi: 10.3389/fpsyg.2016.00765

Howe, E. G. (2010). Sliding "off" the sliding scale: allowing hope, determining capacity, and providing meaning when an illness is becoming worse but a treatment may help. *J. Clin. Ethics.* 21, 91–100.

Kim, S. (2010). *Evaluation of Capacity to Consent to Treatment and Research*. New York, NY: Oxford University Press.

Kim, S. Y., Caine, E. D., Swan, J. G., and Appelbaum, P. S. (2006). Do clinicians follow a risk-sensitive model of capacity-determination? An experimental video survey. *Psychosomatics* 47, 325–329. doi: 10.1176/appi.psy.47.4.325

Kloezen, S., Fitten, L. J., and Steinberg, A. (1988). Assessment of treatment decision-making capacity in a medically ill patient. *J. Am. Geriatr. Soc.* 36, 1055–1058. doi: 10.1111/j.1532-5415.1988.tb04376.x

Lawlor, R. (2016). Cake or death? Ending confusions about asymmetries between consent and refusal. *J. Med. Ethics* 42, 748–754. doi: 10.1136/medethics-2016-103647

Lepping, P., and Raveesh, B. N. (2014). Overvaluing autonomous decision-making. *Br. J. Psychiatry* 204, 1–2. doi: 10.1192/bjp.bp.113.129833

Maclean, A. (2000). Now you see it, now you don't; consent and the legal protection of autonomy. *J. Appl. Philos.* 17, 277–288. doi: 10.1111/1468-5930.00162

Manson, N. C. (2015). Transitional paternalism: how shared normative powers give rise to the asymmetry of adolescent consent and refusal. *Bioethics* 29, 66–73. doi: 10.1111/bioe.12086

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372, n71. doi: 10.1136/bmj.n71

Parker, M. (2004). Judging capacity: paternalism and the risk-related standard. *J. Law Med.* 11, 482–491.

Parker, M. (2006). Competence by consequence: ambiguity and incoherence in the law. *Med Law*. 2006/05/10 edn, 25(1), 1–12.

Re T (Adult: refusal of medical treatment) (1992). 4 All ER 649 at 662.

Roberts, J. T. F. (2018). Autonomy, competence and non-interference. *HEC Forum* 30, 235–252. doi: 10.1007/s10730-017-9344-1

Saks, E. R. (1991). Competency to refuse treatment. *North Carol. Law Rev.* 69, 945–999.

Saks, E. R. (1999). Competency to decide on treatment and research: macarthur and beyond. *J. Contemp. Legal Issues* 10, 103–129.

Saks, E. R., and Jeste, D. V. (2006). Capacity to consent to or refuse treatment and/or research: theoretical considerations. *Behav. Sci. Law* 24, 411–429. doi: 10.1002/bsl.708

Schopp, R. F. (1994). Self-determination and well-being as moral priorities in health care and in rules of law. *Public Aff. Q.* 8, 67–84.

Skene, L. (1991). Risk-related standard inevitable in assessing competence. *Bioethics* 5, 113–117. doi: 10.1111/j.1467-8519.1991.tb00152.x

White, B. C. (1994). *Competence to Consent*. Washington, DC: Georgetown University Press.

Wicclair, M. R. (1991a). A response to brock and skene. *Bioethics* 5, 118–122. doi: 10.1111/j.1467-8519.1991.tb00153.x

Wicclair, M. R. (1991b). Patient decision-making capacity and risk. *Bioethics* 5, 91–104. doi: 10.1111/j.1467-8519.1991.tb00150.x

Wicclair, M. R. (1999). The continuing debate over risk-related standards of competence. *Bioethics* 13, 149–153. doi: 10.1111/1467-8519.00138

Wilks, I. (1997). The debate over risk-related standards of competence. *Bioethics* 11, 413–426. doi: 10.1111/1467-8519.00081

Wilks, I. (1999). Asymmetrical competence. *Bioethics* 13, 154–159. doi: 10.1111/1467-8519.00139

Winick, B. J. (1991). Competency to consent to treatment: the distinction between assent and objection. *Houst Law Rev*. 28, 15–61.

Author Disclaimer: The ideas and opinions expressed are the authors'; they do not represent any position or policy of the National Institute of Health, the Department of Health and Human Services, or the U.S. government.

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 conflictof interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Berens and Kim. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Psychology

TYPE Conceptual Analysis
PUBLISHED 22 July 2022
DOI 10.3389/fpsyg.2022.903960

OPEN ACCESS

EDITED BY
Tomer Fekete,
Ben-Gurion University of the Negev,
Israel

REVIEWED BY
Gualtiero Piccinini,
University of Missouri–St. Louis,
United States
Vanessa Lux,
Ruhr University Bochum, Germany

*CORRESPONDENCE
Matthieu M. de Wit
matthieudewit@muhlenberg.edu
Heath E. Matheson
Heath.matheson@unbc.ca

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Theoretical and Philosophical Psychology,
a section of the journal Frontiers in Psychology

RECEIVED 24 March 2022
ACCEPTED 27 June 2022
PUBLISHED 22 July 2022

CITATION
de Wit MM and Matheson HE (2022) Context-sensitive computational mechanistic explanation in cognitive neuroscience.
Front. Psychol. 13:903960.
doi: 10.3389/fpsyg.2022.903960

COPYRIGHT
© 2022 de Wit and Matheson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Context-sensitive computational mechanistic explanation in cognitive neuroscience

Matthieu M. de Wit[1]*† and Heath E. Matheson[2]*†

[1]Department of Neuroscience, Muhlenberg College, Allentown, PA, United States, [2]Department of Psychology, University of Northern British Columbia, Prince George, BC, Canada

Mainstream cognitive neuroscience aims to build mechanistic explanations of behavior by mapping abilities described at the organismal level *via* the subpersonal level of computation onto specific brain networks. We provide an integrative review of these commitments and their mismatch with empirical research findings. Context-dependent neural tuning, neural reuse, degeneracy, plasticity, functional recovery, and the neural correlates of enculturated skills each show that there is a lack of stable mappings between organismal, computational, and neural levels of analysis. We furthermore highlight recent research suggesting that task context at the organismal level determines the dynamic parcellation of functional components at the neural level. Such instability prevents the establishment of specific computational descriptions of neural function, which remains a central goal of many brain mappers – including those who are sympathetic to the notion of many-to-many mappings between organismal and neural levels. This between-level instability presents a deep epistemological challenge and requires a reorientation of methodological and theoretical commitments within cognitive neuroscience. We demonstrate the need for change to brain mapping efforts in the face of instability if cognitive neuroscience is to maintain its central goal of constructing computational mechanistic explanations of behavior; we show that such explanations must be contextual at all levels.

KEYWORDS

cognitive neuroscience, degeneracy, functional brain mapping, levels of analysis, many-to-many mappings, mechanistic explanation, neural reuse, structure-function relationships

Introduction

A brief history of the origins of modern functional brain mapping

The historical neuropsychological literature is full of case-studies of individuals with specific behavioral impairments following damage to local regions of the cerebral cortex. Perhaps most famously, Broca (1861) described a patient who, after damage to the left inferior frontal gyrus (IFG), permanently lost almost all of his expressive language abilities while his receptive abilities remained largely intact. Other patients may exhibit fluent (but incoherent) expressive language paired with receptive impairments, typically following lesions in the left superior temporal gyrus (STG). From these observations, Wernicke (1874/1969) developed one of the earliest neurocognitive models of language, which centered on the specific functions of IFG and STG and their connection *via* association fibers. Wernicke argued that complex functions such as language are the result of the interaction of multiple, simpler, sensory, motor, and association processes, thereby effectively ending the search for localized higher-level "faculties" in the brain that had dominated the study of brain-behavior relationships up to that time (Fancher, 1996; Bergeron, 2007). Since then, researchers have continued to grapple with understanding the functions of brain parts and their relationship to behavior, and one of the main goals of cognitive neuroscience remains mapping functional descriptions onto neural activity. But what does successful mapping look like?

Functional brain mapping as the analysis, decomposition, and localization of function

The goal of mainstream cognitive neuroscience[1] is to "investigate brain-behavior interactions" and to "address both descriptions of function and underlying brain events" (Journal of Cognitive Neuroscience, 2021); it seeks "an understanding how the functions of the physical brain can yield the thoughts, ideas, and beliefs" of the mind (Gazzaniga et al., 2019, p. 4). In the mainstream cognitive neuroscience research literature, the term *function* is critical to the enterprise. Importantly, function is often implicitly and interchangeably used at multiple, typically three, distinct levels of analysis (Marr, 1982/2010; Craver, 2014; Krakauer et al., 2017; Zednik, 2018; see Garson, 2016, for discussion of conceptualizations of biological function). The first level concerns that of observable behavior or its disorders. Here, cognitive neuroscientists distinguish between various complex behavioral or cognitive categories described at the personal or organismal level; that is, at the level of the behaving human being. For example, we talk about the abilities of (or impairments of) "paying attention," "using a tool," or "speaking." These are the phenomena of which the discipline ultimately seeks a *mechanistic* explanation, in terms of a description of the parts and interactions of a system that gives rise to the phenomenon (Miłkowski, 2013; Craver and Tabery, 2015; Zednik, 2019).[2]

To this end, at the second, subpersonal, level the phenomenon described at the first level is decomposed into components that perform specific computational operations over representations and are organized together in a specific way (Bechtel and Abrahamsen, 2010). These components are therefore identified by their functions (Piccinini and Craver, 2011), the descriptions of which tend to be "human-interpretable" (Hasson et al., 2020). For instance, computational processes are hypothesized that implement, e.g., "attentional orienting," "action selection," or "linguistic retrieval," which are *subcapacities* of the organismal-level capacity of paying attention, using a tool, and speaking, respectively. In other words, and following Wernicke's lead, at the second level multiple latent, interacting, domain-specific or domain-general, functional components are postulated that explain the production of a particular complex behavior or cognitive ability observed at the first level. In general terms, it is widely accepted that computational operations are thought to involve inputs which feed into the manipulation of internal, typically – but not necessarily (Piccinini and Scarantino, 2011) – representational states to produce outputs that are then used downstream in further computational processes (Shea, 2018).

The third level at which function is invoked is that of the brain. Thus, following the characterization of a componential computational architecture at the second level (and sometimes *before* this characterization has taken place; see Agis and Hillis, 2017), an attempt is made to map the components onto the activity of the brain, effectively reifying that architecture; that is, the researcher tries to describe the coordinated computational operations in terms of physical neural processes (Piccinini and Shagrir, 2014; Burnston, 2021). There are at least two possible characterizations of this final step. Sometimes, the assumption is that a computational operation will directly map onto a specific brain "part," where a part is broadly construed as a cell, localized assembly of cells or a distributed

[1] While cognitive neuroscience is currently dominated by cognitivist assumptions of representation, computation and mechanism (which we here call the mainstream approach), 4E (embodied, embedded, enactive, extended) as well as ecological approaches to the study of the brain are becoming increasingly prevalent. Further, recent work aims to merge key insights from both traditions (e.g., Piccinini, 2022). In the present article, we have chosen to limit our analysis of research practices as they occur in the mainstream cognitive neuroscience literature. A similar analysis of neuroscientific research practices within 4E and ecological frameworks would be a rather different enterprise and is outside the scope of the current article (for examples of work in this tradition, see Gibson, 1966; Chiel and Beer, 1997; Barrett, 2011; van Orden et al., 2012; Dotov, 2014; Anderson, 2014; Kiverstein and Miller, 2015; de Wit et al., 2017; Hutto et al., 2017; Dewhurst, 2018; Bruineberg and Rietveld, 2019; de Wit and Withagen, 2019; van der Weel et al., 2019; Ryan and Gallagher, 2020; Raja, 2021; Raja and Anderson, 2021).

[2] The topic of how cognitive neuroscience has come to decide on which explananda are worth pursuing is interesting on its own, and a point we return to later, in the section "Discussion: Implications for Cognitive Neuroscience" (see Fancher, 1996; Kästner, 2017 for discussion). Further, more detailed discussion of the role of mechanism in science is beyond the scope of this manuscript (see Cartwright et al., 2020). Our argument assumes a commitment to understanding cognition mechanistically, that is, being explicable in terms of the joint action of parts. Even if one denies a specific characterization of mechanism as discussed in the philosophy of science, we think it uncontroversial that there is at least an implicit and informal understanding that cognitive neuroscience, like the cognitive psychology it inherits, is often seeking mechanistic explanation.

network of cells (Glasser et al., 2016). For instance, using lesion-symptom mapping or functional magnetic resonance imaging (fMRI), cognitive neuroscientists may search for a neural region or network of regions whose function it is to implement the computations that instantiate attentional orienting (while sometimes considering the biological constraints of the part such as, e.g., neural response latencies or receptor types; e.g., Kravitz et al., 2013), and another region or network whose function it is to implement the computations for action selection. However, sometimes the assumption is that a computational operation requires additional decomposition before it can be mapped onto the brain. For instance, using computational modeling of different types of cells or cell networks, attentional orienting is broken down further into component operations (e.g., a number of different computations in a connectionist model). Either way, the goal is to mechanistically explain organismal-level phenomena, *via* organized, interacting computational components, in terms of functionally specialized brain cells, assemblies, or networks that combine into interacting brain parts.[3] Though cognitive neuroscience has moved on from structure-function mapping understood in the modular sense described earlier (e.g., expressive language is due to Broca's area), many cognitive neuroscientists – implicitly or explicitly – still seek to discover *the* brain basis of our ability to pay attention, use a tool, or speak (Cabeza and Nyberg, 2000; see Zerilli, 2019 for a historical overview of this approach).

Given this, brain mapping continues to be a major effort, which can also be gleaned from its influence on the infrastructure of the field (e.g., the UCLA Brain Mapping Center; the journal *Human Brain Mapping*). Indeed, the idea of functional brain mapping is so deeply engrained in current thinking about brain-behavior relationships that it is implicit in our nomenclature and pervades textbooks and day-to-day public discourse about neuroscience. Thus, there is talk of the primary *visual* cortex or the dorsal *attention* network, which are believed to implement specialized computations underlying visual and attentional behavior, respectively. In a sense these efforts are clearly successful. For instance, it is possible to manipulate the activity of cortical regions using transcranial magnetic stimulation (TMS) and observe predictable effects on behavior, lesions to early visual regions result in predictable visual field deficits, and sophisticated computational models exist that impressively replicate various properties of real-world brain networks.

However, at the same time, the practice of functional brain mapping as defined above has met serious empirical and theoretical challenges, particularly in recent years (Uttal, 2001; Pessoa, 2008; Anderson's, 2010, Anderson, 2014; Klein, 2012; Burnston, 2016b; Khalidi, 2017; Stanley et al., 2019; Viola, 2020; Zerilli, 2020). We are certainly not the first to note challenges in this domain: Previous neuroscientists critical of the functional brain mapping enterprise have called for a more effective integration of research findings at each level of analysis (Krakauer et al., 2017), have argued for a redescription of function at the level of local brain regions (Price and Friston, 2005; Poldrack, 2010), or have pointed out general limitations of current mapping efforts (Poldrack, 2006; Genon et al., 2018), among other criticisms. However, with some important exceptions to be discussed below, several of these researchers still aim to identify *the* componential computational operations of well-defined brain parts and in that sense maintain the central premises of functional brain mapping. For example, Shine et al. (2016, p. 26) state that "[i]n considering the computational capacities of independent brain regions, we will make the argument that computational specialization is not only abundant in the brain, but also that it would be difficult to imagine a working brain that did not contain such specialization." However, the viability of this approach is in question.

Outline of the review and further analysis

In contrast to previous critiques, we will suggest that the core problem cognitive neuroscience faces is an epistemological one: The present paper will integrate empirical and philosophical literature to show that the goal of giving *any* specific, computational description of *context-independently* defined brain parts is not possible, and therefore that the explanatory strategy of mainstream cognitive neuroscience is in need of revision. The section "Challenges to the Practice of Structure-Function Brain Mapping: An Integrative Review" provides an integrative review of the empirical literature on flexible neural tuning (section "Neural Tuning and Functional Brain Mapping"), plasticity and recovery of organismal-level function following brain lesions (section "Lesion Studies, Plasticity, and Functional Brain Mapping"), inter and intra-individual neural degeneracy (section "Degeneracy and Functional Brain Mapping"), neural reuse (section "Neural Reuse and Functional Brain Mapping"), and the neuroscience of enculturated skills such as reading and mathematics (section "Enculturated Skills and Functional Brain Mapping"). These areas are increasingly being investigated by researchers within cognitive neuroscience, and we summarize one consequence of these challenges that has been recognized in the field as "weak contextualism," which denotes variable

[3] Note that different researchers may flesh out the second, computational level to a greater or lesser degree, and describe its connections to the third, neural level in more or less detail. Some researchers build full-fledged sophisticated computational models without paying much attention to the neural level, while other researchers attempt to characterize functions of brain regions using a combination of behavioral and neuroimaging methods, without fleshing out the computational operations of those regions in any detail.

mappings relative to the organismal level of description (section "Consequences of Weak Contextualism"). However, section "Strong Contextualism, Instability of Mapping, and Indeterminate Part Ontology for Cognitive Neuroscience" describes the fundamental incommensurability of the goals of mechanistic explanation and any context-independent – even weak contextual – descriptions of brain function, and highlights a consequence not yet widely recognized within the field, a type of instability in functional mapping that reflects a "strong contextualism." Most significantly, we review recent research that forces a reconsideration of what constitutes a relevant neural "part" to begin with and show that the parcellation of functional components shifts depending on the task context we choose to study at the organismal level. Finally, the section "Discussion: Implications for Cognitive Neuroscience" provides a brief methodological and theoretical sketch of a cognitive neuroscience that can maintain its central goal of constructing robust computational mechanistic explanations of behavior by being sensitive to the fact that such explanations must be contextual at all levels.

Challenges to the practice of structure-function brain mapping: An integrative review

Below we provide an integrated review of research that present challenges to structure-function mapping (see Ames and Fiske, 2010; Anderson, 2014; Seifert et al., 2016; Hartwigsen, 2018; Rule et al., 2019 for previous focused reviews of each of the topics discussed in this section). Afterward, we will highlight the consequences this literature has had on the current state of the art in cognitive neuroscience.

Neural tuning and functional brain mapping

Neuroscience has a long history of characterizing the neural tuning of brain regions, where manipulations of stimulus parameters of popular interest (e.g., line orientation, face-ness of an object, etc.) are correlated with changes in neural activity (firing rate, BOLD signal, etc.; see Buzsáki, 2020, for a brief description of this history). However, neuroimaging results show that many neural regions are not statically tuned to particular types of stimuli in a stable manner (Bair, 2005; Clopath et al., 2017). While flexibility in neural tuning at the single neuron level has been shown for some time (e.g., Miller, 2000), recent results show that the neural tuning of most brain regions appears capable of changing rapidly between different cognitive tasks. For instance, Çukur et al. (2013) had participants watch videos while lying in the scanner. They were instructed to attend to different features of the videos, specifically vehicles or humans. As they did so, it was shown that the tuning characteristics of almost every region in cortex shifted depending on the goal of the observer (i.e., voxel activity was better explained by responding to humans when searching for humans and vice versa for vehicles). Other research has demonstrated more specific effects. For instance, ventral cortical regions, typically implicated in identifying objects, do not do so in an all-or-none fashion but shift their response tuning to object identity depending on the exact task participants are performing (Harel et al., 2014). In addition, recent advancements in multi-voxel pattern analysis (MVPA; Kriegeskorte et al., 2008) suggest that the "representational geometry" (i.e., the abstract, multidimensional space of neural activity patterns; see Kriegeskorte and Diedrichsen, 2019) of different network nodes is dynamically defined by context. For instance, geometry of areas in the dorsal stream is better described by action similarity in a task where participants make judgments about action, but category similarity when participants make judgments about category (see also Anderson and Oates, 2010; Gallivan and Culham, 2015; Bracci et al., 2017). Similarly, patterns of activity in some areas of the cortex are implicitly tuned to the category of objects depending on cues that are available in the environment (Matheson et al., 2021). The important thing to note here is that the tuning of single cells determines the representational geometry that is measured in these studies (Kriegeskorte and Wei, 2021). Thus, changes in neural representational geometry in different tasks reflect changes in neural tuning across most of the cortex depending on the context.

Overall, these neural tuning findings challenge functional mapping efforts because they suggest that the target – the response properties of a neural region to particular stimulus information – is a moving one that is shaped by task context.

Lesion studies, plasticity, and functional brain mapping

Second, findings of neural plasticity and recovery have long complicated the functional brain mapping literature. After (sometimes extensive, e.g., García et al., 2017; Bowren et al., 2021) brain damage, patients may be able to partially or fully recover the behavioral or cognitive ability that was lost (Kolb and Gibb, 2013). This trajectory of recovery can continue for years (Hartwigsen and Saur, 2019) and has even been observed following lesions in early sensory areas (in adult cats; see Jiang et al., 2021). One account of this phenomenon, consistent with the assumption that local neural regions perform specialized computational operations, is that the regained organismal-level function is supported by different computational processes. Perhaps a compensatory mechanism, involving different cognitive strategies, restores function at the organismal level (Dixon et al., 2008; see De Brigard, 2017 for a related discussion focusing on brain and cognitive strategy changes associated with healthy aging).

Another more commonly offered account of recovery that is consistent with functional brain mapping is that, as a result of plasticity, the remaining undamaged brain regions are able to reorganize themselves to accommodate new functions; that is, neural parts are repurposed to perform new computational operations. However, if recovery is indeed a matter of repurposing (i.e., the specific computational operation formerly realized by the damaged area is now performed by another area), one would expect to lose the function that the newly colonized area was responsible for. Alternatively, it is commonly postulated that the remaining intact tissue is utilized more efficiently after recovery, allowing for the former and the new computational operation to be implemented alongside each other, but this notion of redundancy (Friston and Price, 2003) begs the question why the colonized tissue was utilized less efficiently before, given that neural tissue is notoriously expensive to maintain. To our knowledge, observations of loss of one function accompanying recovery of another function appear to be largely absent from the patient literature.

On the contrary, several recent treatment studies have reported gains in the *language* domain as a result of *upper extremity* movement therapy in stroke patients (Harnish et al., 2014; Primaßin et al., 2015; see Anderlini et al., 2019 for review; and Stoll et al., 2021 for related work in limb apraxia). Furthermore, brain damage is often associated with multiple co-occurring deficits (e.g., patients presenting with both limb apraxia and aphasia following left hemisphere lesions), similarly suggesting that the impacted area supports diverse behavioral domains (Behrmann and Plaut, 2014; Goldenberg and Randerath, 2015). So called "crossed" cases of classical clusters of deficits have moreover been reported in patients with atypical lateralization of function (e.g., limb apraxia presenting together with aphasia following a *right* hemisphere lesion; Raymer, 1999; see Vingerhoets et al., 2013 for findings of co-lateralization in neurologically intact individuals). Each of these findings is hard to explain under the assumption that impacted functions are implemented by specialized neural parts, and may be interpreted as further evidence that the functional significance of any given region is sort of a chameleon which does not permit a context-free definition (Price et al., 2016; Price, 2018).[4]

Degeneracy and functional brain mapping

It is increasingly recognized that the phenomenon of neural degeneracy – the notion that structurally different neural processes can produce equivalent behaviors at the organismal level – plays an important role in the brain (Edelman and Gally, 2001; Price and Friston, 2002; Noppeney et al., 2004, 2006; Figdor, 2010; Bateson and Gluckman, 2011; Sporns, 2011; Marder et al., 2015; Anderson, 2016; Seifert et al., 2016; De Brigard, 2017; Viola, 2020). Both inter- and intra-individual cases of degeneracy have been observed (Anderson, 2016). In the domain of numerical cognition, Tang et al. (2006) reported markedly different patterns of activation during simple arithmetic between native Chinese and native English speaking individuals, with left perisylvian activation in the former cultural group, and a network of "visual" and "premotor" regions in the latter, despite equivalent stimuli (Hindu-Arabic numerals) and equivalent performance at the behavioral level. In the language domain, Biduła et al. (2017) reported many different variants and degrees of language lateralization in neurologically intact individuals with normal language abilities, involving, in addition to more or less typical and atypical lateralization of classic language areas, right hemisphere components of the "default mode network" as well as an atypical role for the cerebellum. Finally, Merabet et al. (2008) provide evidence for intra-individual degeneracy in neurologically intact individuals by showing the existence of multiple, different neural substrates for braille reading (see De Brigard, 2017 for additional examples related to healthy aging). Degeneracy is clearly widespread, both in the intact and the lesioned brain (Fotopoulou, 2014; Price, 2018).[5] Indeed, we would argue that functional recovery following brain damage can be considered a prime example of both inter and intra-individual degeneracy (see also Mogensen, 2011; Abrevaya et al., 2017; Hartwigsen and Saur, 2019).

Neural reuse and functional brain mapping

Fourth, there is extensive evidence from neurologically intact individuals showing that some, if not most, brain regions are implicated in many different behaviors, suggesting that they can be reused in different contexts. That is, regions are typically capable of participating in a diverse array of functions. The above-mentioned study by Merabet et al. (2008)

[4] A common alternative interpretation of the co-occurrence of symptoms is that lesions "do not color within the lines." That is, that they may impact multiple – smaller – spatially co-located functionally specialized regions. However, this line of reasoning does not seem readily capable of accommodating the above-reported "crossed" cases and the impact of treatment in one behavioral domain on performance in another behavioral domain.

[5] It is important to exclude the possibility that what seems like degeneracy at the neural level is in fact driven by organismal-level variability between or even within participants in cognitive or behavioral strategies when performing a task (Gardner et al., 2013; Berneiser et al., 2018).

showed, using TMS, that the occipital (i.e., "visual") cortex of blindfolded sighted participants became causally involved in tactile perception following an intense 5-day braille reading training program, providing evidence for its functional perceptual capacity beyond the visual modality (see also Murray et al., 2015). Similarly, in congenitally blind individuals who have never possessed sight, the occipital cortex has been shown to be sensitive to non-perceptual stimulus attributes such as the grammatical structure of spoken sentences and the difficulty of math equations (Bedny, 2017). Much evidence has been marshaled in the last decade or so showing that even gross functional distinctions at the organismal level, for instance the difference between emotional processes and cognitive ones (Pessoa, 2008) or between perception and action (Cisek, 2007), do not hold at the neural level, due (in part) to reuse.

The consistency of reports of functional heterogeneity suggests that reuse is not a curiosity but a general feature of the nervous system (Anderson, 2010, 2014). Importantly, despite empirical advancements showing that some regions can be functionally further subdivided (e.g., Broca's area; Fedorenko and Blank, 2020) reuse is observed regardless of the level of granularity at which these analyses are performed (i.e., whether the brain was parsed into, say, 10 or a 1,000 regions); thus, reuse may be reduced (Poldrack, 2006) but does not go away at an increased spatial resolution (Anderson et al., 2013; Uddin et al., 2014). For instance, it is clear that, at the resolution of the entire brain, the entire brain is reused to support different behaviors, but neural reuse can be observed even at the resolution of single neurons, with neurons involved in either sensory or motoric functions depending on the behavioral and concomitant neural context in the roundworm *C. elegans* (Bargmann, 2012). The fact that reuse phenomena do not disappear at smaller resolutions presents a major challenge to determining structure-function mappings.

Enculturated skills and functional brain mapping

Fifth and finally, Dehaene and Cohen (2007); (see Menary, 2015; Jones, 2018 for discussion) suggest that cognitive tasks that require enculturation and formal schooling in order to be displayed (like reading, writing, and mathematics) are supported by the neuronal "recycling" of neural regions that originally evolved for other purposes, yet have the right structure to implement those tasks – again a type of reuse, though reuse in this case is defined at longer timescales (see Borra and Luppino, 2018 for additional examples). The visual word form area in occipitotemporal cortex is a good example of such a region. The recycling account is convincing because these abilities have arguably emerged too recently (i.e., within the last few thousand years) for evolution to have generated specialized cortical regions to support them.

Consequences of weak contextualism

This brief review integrates some of the most significant challenges facing cognitive neuroscience. Within the field, it is increasingly recognized that all of these phenomena suggest that the functional role of a brain region is context-dependent. However, while context-dependence is recognized by the field, we argue that this is recognition of a type of "weak" contextualism. By weak contextualism we mean that most researchers accept that behavioral context shapes a region/network's functional role in organizing the organismal-level behavioral phenomenon, but still maintain that the functioning of the part itself is context-independent. That is, it is thought that *functions of brain parts are not stable when defined relative to the organismal level, but functions defined at the computational and neural level are* (compare to Burnston's, 2016a, 2019 "absolutism"). Researchers sympathize with weak contextualism when they make continued calls for context-independent computational descriptions of brain regions while recognizing that this computation implements a cognitive subcapacity that contributes to many different organismal-level behaviors. For example, Vingerhoets et al. (2013) report evidence in neurologically intact individuals that skilled action and expressive language involve strongly overlapping neural components. To account for this overlap, they postulate that these components implement the production of complex (i.e., precise, articulated, coordinated, nested) learned movement, an operation that is common to both tool use and speech, explaining its involvement in each of those contexts (see Knops et al., 2009; de Wit et al., 2012; Parkinson et al., 2014 for similar conceptualizations of shared informational or computational demands across different behavioral domains). Thus, with this approach, the functional description of neural parts is context-independent and is abstract enough to account for their contextual involvement in a wide range of tasks (Price and Friston, 2005; Shine et al., 2016; Humphreys et al., 2021; see also Anderson's, 2010 early "working" vs. "use" conceptualization of neural reuse). Note that this type of contextualism, though now often accepted in the field, is already quite far removed from the traditional structure-function accounts described in the section "Introduction," which have typically characterized the function of brain regions relative to the organismal level (e.g., the fusiform face area is important for face identification) or at a minimum in terms of cognitive subcapacities directly related to specific organismal-level abilities (e.g., a region involved in "attentional selection"). Regardless, an acceptance of weak contextualism would still allow us to find *the* computational function of a brain part (as was suggested in the case of Vingerhoets et al., 2013).

Strong contextualism, instability of mapping, and indeterminate part ontology for cognitive neuroscience

While there may be sympathies toward weak contextualism within cognitive neuroscience, we argue that there is, inescapably, a form of "strong" contextualism that is not widely acknowledged (and therefore one that is far from accepted). Our argument builds on arguments within the philosophy of science regarding the consequences of seeking mechanistic explanation. Again, by mechanistic explanation, we mean the goal of providing a description of the parts and interactions of a system that give rise to a phenomenon (e.g., Craver, 2014), and in cognitive neuroscience the typical approach is to seek a mapping of behavior to computation to brain. Indeed, some philosophers have argued that a type of strong contextualism is an unavoidable consequence of seeking mechanistic explanations in general (i.e., not just a problem for cognitive neuroscience; see Lee and Dewhurst, 2021 for discussion) and therefore it is an unavoidable consequence of the explanatory goals of cognitive neuroscience. Here, we highlight two critical arguments to demonstrate strong contextualism; one regarding the functional mapping of the computational level to the brain level, and one regarding the parcellation of brain parts. We hope to show that these two epistemological issues demand methodological and theoretical re-orientation within the field that is much more significant than the demands of weak contextualism.

Dynamic functional mappings

First, our integrative review leads to the conclusion that behavioral context not only determines the contribution of a brain part to the organization of behavior (i.e., weak contextualism), but that *context determines which computational operation that brain part implements* in support of the behavior. That is, the computation a region performs is not a specialization *of that region*, but rather is determined by the behavioral and neural context in which the region finds itself, and can shift when the context changes (Sporns, 2011; Klein, 2012; Anderson, 2014, 2015a; Burnston, 2016b, 2019, 2021; Khalidi, 2017; and see Mesulam, 1990; McIntosh, 1999, 2000, 2004 for early arguments in this direction). Klein (2012) illustrates the idea clearly with a discussion of the function of pistons in trucks with engine brakes: "Most of the time, [pistons] compress a fuel-air mixture to the point of detonation and transmit the generated power to the crankshaft. On trucks equipped with engine brakes, the pistons also have a second function: when the engine brake is engaged, the pistons use power from the wheels to compress air in the cylinder, slowing the truck. *Which function the piston performs depends on things external to it*: whether it is powering or slowing the truck depends on the ignition system and the valve timing" (p. 955, italics added; in this metaphor, this is the neural context). Notice that the function of the piston depends on whether we are interested in explaining the "going" or "stopping" action of the truck (the behavioral context); specifically, it is causally transmitting explosive force to the crankshaft in one instance and is reacting to a vacuum in the cylinder in the other. Here, we have one part (the piston) that is not simply performing an abstract function useful to both stopping and going (cf. Vingerhoets et al., 2013), but that has a functional description that is dependent on our explanatory goals. Thus, under the strong contextualism view, there is nothing specialized about the functional role of a piston – there is *no* single computational operation performed by the part that plays a role across phenomena. Thus, the mapping of the part to the computation is unstable. We argue that this conclusion holds for brain parts. The empirical evidence collected in the last section is consistent with the idea that, when seeking mechanistic explanations, brain regions are best modeled with different computations in different tasks (regardless of whether the computations are described mathematically or verbally). For example, in one context a brain region's activity might be best modeled as multiplying an input signal whereas in another it is best modeled with addition; in one context a network might be best described as an "integrator," while in another it is a "filter," etc. Thus, *functions of brain parts are not stable regardless of whether they are defined relative to the organismal, or computational, or neural level.*

Note that we are not denying that neurons (and neural networks) have physiological, morphological and other neuroanatomical (e.g., topological) characteristics that ensure they can do some things and not others, in the same way we wouldn't deny a piston's physical characteristics that allow it to do some things and not others. Our point is that a structure's participation in any given behavior, while obviously – and importantly – constrained by its properties, is not determined by those properties, and we cannot describe a part's functional contribution in a context-independent way *at any level* within our mechanistic explanation.

Importantly, not even extreme abstraction of the putative computational function will allow us to recover a stable structure-function mapping for a part. This is the case given that we can choose to study an infinite number of phenomena at the behavioral level and that we will evolve skills in the future that we have not characterized yet. For instance, while the "production of complex learned movement" attributed to a neural network may apply to explanations of skilled action and expressive language (Vingerhoets et al., 2013), we are unable to rule out that the same neural network supports some behavior that does not require "complex learned movement," simply because we haven't studied all existing behavioral phenomena, nor can we know what phenomena will arise in the future. Thus, unlike weak contextualism, strong contextualism reveals an instability in structure-function mapping that prevents us from ever finding *the* computational function of brain parts.

Dynamic part ontologies

The second component of our argument relates to brain parts themselves. Because the identification of parts plays a central role in mechanistic explanation (Kaiser, 2018), identifying brain parts (defined as networks, anatomical regions, subregions, circuits, or single cells) is a cornerstone of cognitive neuroscience. Thus, a *central* issue in a cognitive neuroscience that takes mechanistic explanation seriously is determining the right "part ontology" (cf. Stanley et al., 2013; Viola and Zanin, 2017). However, strong contextualism shows that context determines not only a part's computational role in any given behavior, but also determines what we should even consider to be a relevant part in the first place – that is, *context determines the appropriate ontology of parts* (cf. Poldrack and Yarkoni, 2016; Genon et al., 2018; Uddin et al., 2019). Strong contextualism challenges the idea that there is any context-independent parcellation of the brain that cognitive neuroscience can use in its mapping efforts. That is, changing our explanatory goals (i.e., which behavioral phenomenon we seek to explain) results in a redefining of the causally relevant parts that give rise to the phenomenon (see Craver and Kaplan, 2020). Indeed, it follows from the strong contextualism described here that (1) the boundaries of functionally relevant brain parts can shift every time we identify and want to explain a new behavioral phenomenon of interest and map the requisite computations onto brain parts, (2) these boundaries need not follow any obvious structural boundaries, and (3) they might shift even within participants, depending on context. These are not mere speculations. Consistent with point (2), functional boundaries within the brain need not follow any obvious neuroanatomical boundaries. For example, King et al. (2019) reported that functional subdivisions of the cerebellum did not coincide with lobular boundaries. Further, parcellation requires thresholding and clustering approaches, and it is well-known that different approaches will lead to different neuroanatomical maps (Sporns, 2012). Additionally, if, as we have argued above, the computational operation of a part is not context-independently predetermined by its material properties alone, then there is no reason to assume that its size, shape, or position *would* be context-independently definable for neural explanations of behavior. To see this, consider the following instructive metaphor in which a fictional researcher is interested in mapping function to structure in the extraneural human body. In describing high-fiving, the researcher identifies the hand as a whole as a functional part, while in describing the feeling of soft materials for pleasure, (parts of) individual finger segments constitute the relevant parts. In line with this analysis, and consistent with (1) and (3), Salehi et al. (2020) recently found that the boundaries of functional brain parts shift depending on the behavioral state of the participant. More specifically, they "demonstrate that the parcels are indeed consistent for a given condition, but reproducibly reconfigure across conditions, even when starting with the same initial atlas each time" (p. 2). This evidence shows that the foundational assumptions of cognitive neuroscience that it can map *the* brain parts to the computations that support organismal level behavior are not tenable.

Overall, strong contextualism presents an underappreciated characterization of brain-behavior interactions, revealing that context shifts are associated with instability of mapping between organismal, computational, and neural levels. It also leads to the counterintuitive conclusion that context determines part ontology (and not the reverse; see section "Discussion: Implications for Cognitive Neuroscience" below for more details). This epistemological consideration lays bare a unique challenge to cognitive neuroscience's goal of "an understanding how the functions of the physical brain can yield the thoughts, ideas, and beliefs" of the mind (Gazzaniga et al., 2019, p. 4).

Discussion: Implications for cognitive neuroscience

As discussed above, much of the current infrastructure of cognitive neuroscience has been and continues to be shaped by structure-functional brain mapping, and this has real consequences for how money, time, and physical resources get distributed, which includes how undergraduate and graduate students are introduced to the field and how knowledge is classified and disseminated. In our opinion, strong contextualism should force the field to seriously reconsider its approach to the study of how organismal-level functioning maps onto the computational level and how this maps onto the neural level. Below, we gesture toward a number of implications that researchers should consider when pursuing functional brain mapping.

Strong contextualism forces a shift in the goal of mainstream cognitive neuroscience toward developing relevant part ontologies and seeking computational mechanistic explanation *in context*. First, we must recognize that by deciding that we are interested in a particular organismal-level phenomenon in a particular situation (say, visual decision making in a dual rather than a single task setting), we "fix" the part ontology which best helps explain the behavior computationally – though of course we don't know in advance what that part ontology is (in the same way that deciding whether we are interested in going or stopping helps fix the part ontology for the best explanation of truck behavior). Thus, cognitive neuroscience should abandon the aim of describing the context-independent specialized functions of well-defined brain parts. Rather, in specific contexts it should seek to answer the empirical question of what are *relatively stable parcellations*, likely at the level of the individual or at a minimum of a small subgroup of individuals, that together with context-dependent computational descriptions lead to robust explanations of the behavioral phenomena of interest (Salehi et al., 2020; Viola, 2020). That is, we switch from asking "What does neural part X do, and is it important for behavior Y?" to "Given behavior Y, what neural parts – maybe X, maybe Z – and associated putative computational functions are needed to explain it?"

Second, contrary to previous proposals that attempt to deal with weak contextualism (Price and Friston, 2005; Poldrack, 2010), our analysis suggests that the determination of a part ontology should actually take place at a relatively *low* (rather than high) level of abstraction, rendering it more useful for explanations in specific contexts (Klein, 2012; Anderson, 2014; Burnston, 2016b, 2021). For instance, the hippocampus is a behaviorally promiscuous region that appears in explanations within many areas of cognitive neuroscience, most commonly in accounts of spatial or mnemonic behaviors (Jeffery, 2018, for a brief history) but other types of tasks as well, leading some to suggest there is an impasse in theorizing about hippocampus function (Ekstrom and Ranganath, 2018; see Humphreys et al., 2021 for a similar discussion centering on the angular gyrus). Consistent with weak contextualism, one approach to this staggering complexity has been to offer relatively abstract functional descriptions that could account for the hippocampus's role in all of these phenomena, for instance pattern separation/completion (Yassa and Stark, 2011), scene construction (see Maguire et al., 2016 for debate), or context equivalency (Maurer and Nadel, 2021). Notice that these are all descriptions of functions that are abstract enough to be equally applicable across many parts of the brain and clearly show conceptual overlap (in at least the verbal accounts of their computational roles); further, it is hard to imagine identifying a behavioral phenomenon that would not implicate these types of computational functions. In contrast, strong contextualism points to why such functional promiscuity exists, and suggests that the hippocampus may be best understood, mechanistically, as a part that implements one type of computational operation that supports some behaviors in some tasks (e.g., memory in memory tasks), but other operations in other tasks (e.g., in navigation tasks). Importantly, the hippocampus, as a brain part defined in a context-independent way, may not be the right "part" for explanations of some other tasks which may require further decomposition into additional parts of the hippocampus itself (e.g., in imagination tasks). Another natural consequence from this analysis is that a functional part may crosscut gross neuroanatomical boundaries. As a hypothetical example, in some contexts but not others subregions of both the angular gyrus and the supramarginal gyrus may together constitute a functional part. Ultimately, the part ontology will depend on whether we are interested in explaining memory, navigation, or imagination.

Note that we are not denying that the hippocampus (and its neural contexts) have anatomical structure and physiological properties that we can characterize – and in fact, in moving away from focusing on context-independent functional descriptions, the description of such constraints takes on an increased importance (see also Anderson, 2015a; Bolt et al., 2017). For instance, there is evidence that, given its physiology, the hippocampus implements sequence generation (Buzsáki and Tingley, 2018). However, the question is about the context-independent relevance of any given anatomical or physiological feature for explanations within cognitive neuroscience. We are suggesting that the degree to which a particular part (e.g., hippocampus) with particular properties will make a stable appearance in computational *mechanistic explanations* of particular behaviors across contexts, and whether the mechanistic explanation (with a particular part ontology) results in strong predictions for behaviors of interest, is an empirical question to be addressed in cognitive neuroscience under the specter of strong contextualism.

Finally, the implications of strong contextualism extend beyond strategies for empirical work; it affects how we theorize and interpret results. Indeed, it is recognized by many that the behavioral and neural sciences are in a "theory crisis" (Eronen and Bringmann, 2021). Our review suggests that one source of the problem is failing to recognize the instability of mapping "paying attention," "using a tool," and "speaking" to computational operations and brain parts. Strong contextualism, then, suggests an additional way to address the theory crisis: By developing formal computational mechanistic explanations in context, cognitive neuroscience can explicitly test effective part ontologies with associated computational operations and functional roles based on the phenomena it seeks to understand, *where the phenomena it seeks to understand are open to revision and new classification*. Doing so opens up cognitive neuroscience not only to better "part ontologies" but also to more useful cognitive ontologies (Poldrack, 2010). Such an approach may lead to further advancements in which historic ontologies may be dismantled in favor of ones with potentially greater accuracy and hence hopefully wider reaching clinical and scientific impact (e.g., see Renoult et al., 2019, for an example of the dismantling of the episodic vs. semantic distinction and challenges this poses to conceptions of "memory"; see Anderson, 2011 for challenges to the reification of "attention"; see Buzsáki, 2020, for challenges to the distinction between perception and action, etc.). Thus, we foresee an iterative process in which respecting the notion that context fixes the part ontology that best explains behavior allows us to remain open to other functional mappings in other contexts, and searching for empirically useful part ontologies will also shape how we identify and characterize the behavioral phenomena we wish to explain (see also Anderson, 2015b). This is a suggestion that goes beyond stating we need to understand behavior better before we should seek mechanistic explanations (Krakauer et al., 2017); rather, it opens up the possibility of characterizing both part and behavioral ontology differently than we do now, making way for new theoretical approaches and insights into brain-behavior interactions in a way that allows cognitive neuroscience to feed back into the behavioral sciences.

We have encountered resistance to the arguments presented here. Strong contextualism is counterintuitive and there are two common reactions to these arguments that we want to briefly address before we conclude. First, one might accept weak contextualism but conclude the strong version is a step too far. Again, as we have shown, many researchers hold that there is a single abstract computational process that each region performs, and we just have to wait for technology and/or methodology to catch up before we can successfully map that region to a computation that maps to behavior. We have already addressed the limitations of this conclusion but want to highlight here that our arguments regarding strong contextualism are not technological or methodological, but primarily epistemological (and our integrative review provides empirical support for the epistemological issues we have highlighted). Advances in technology or methodology without theoretical advances will not address the problems identified in the sections "Challenges to the Practice of Structure-Function Brain Mapping: An Integrative Review" and "Strong Contextualism, Instability of Mapping, and Indeterminate Part Ontology for Cognitive Neuroscience."

Second, one might argue that contextualized explanation is already the current state of the art, in which researchers in disconnected (siloed) subfields of cognitive neuroscience identify context-sensitive functions for brain parts, no one is *really* looking for context-independent explanations in cognitive neuroscience, and everyone already acknowledges that there is no principled way of discovering *the* mapping between a structure and its computations. If this was indeed the state of the art, there would be no reason for the anatomical parcellations to be similar across silos, as accepting strong contextualism means accepting empirically defined parcellations that will be contextually driven. However, parcellations *are* similar across silos (e.g., hippocampus is mainly structurally defined in both memory and navigation research), suggesting that researchers do think there is a meaningful context-independent parcellation of the brain that can be mapped to computations and behaviors of interest.

Ultimately, the efforts of cognitive neuroscience will be judged by their utility in fulfilling its goals as stated above (Gazzaniga et al., 2019; Journal of Cognitive Neuroscience, 2021). It is our contention that taking strong contextualism seriously, both when it comes to determining computational operations, and when it comes to determining the size, shape and location of the neural parts that instantiate these operations, will position us better to mechanistically explain the clinical behavioral observations that this article started with, and even to more fully understand what it means for us to pay attention, use a tool, or speak to one another.

Author contributions

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

Funding

HM acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grant RGPIN-2018-06139.

Acknowledgments

HM would like to thank the students in the lab at UNBC for fruitful discussions of the main thesis.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abrevaya, S., Sedeño, L., Fitipaldi, S., Pineda, D., Lopera, F., Buritica, O., et al. (2017). The road less traveled: alternative pathways for action-verb processing in parkinson's disease. *J. Alzheimers Dis.* 55, 1429–1435. doi: 10.3233/JAD-160737

Agis, D., and Hillis, A. E. (2017). The cart before the horse: when cognitive neuroscience precedes cognitive neuropsychology. *Cogn. Neuropsychol.* 34, 420–429. doi: 10.1080/02643294.2017.1314264

Ames, D. L., and Fiske, S. T. (2010). Cultural neuroscience. *Asian J. Soc. Psychol.* 13, 72–82. doi: 10.1111/j.1467-839X.2010.01301.x

Anderlini, D., Wallis, G., and Marinovic, W. (2019). Language as a predictor of motor recovery: the case for a more global approach to stroke rehabilitation. *Neurorehabil. Neural Repair* 33, 167–178. doi: 10.1177/1545968319829454

Anderson, B. (2011). There is no such thing as attention. *Front. Psychol.* 2:246. doi: 10.3389/fpsyg.2011.00246

Anderson, M. L. (2010). Neural reuse: a fundamental organizational principle of the brain. *Behav. Brain Sci.* 33, 245–66; discussion266–313.

Anderson, M. L. (2014). *After Phrenology: Neural Reuse And The Interactive Brain*. Cambridge, MA: The MIT Press.

Anderson, M. L. (2015a). Beyond componential constitution in the brain: starburst amacrine cells and enabling constraints. *Open MIND* 1, 1–13. doi: 10. 15502/9783958570429

Anderson, M. L. (2015b). Mining the brain for a new taxonomy of the mind. *Philos. Compass* 10, 68–77. doi: 10.1111/phc3.12155

Anderson, M. L. (2016). "Neural reuse and in-principle limitations on reproducibility in cognitive neuroscience," in *Reproducibility: Principles, Problems, Practices, and Prospects*, eds H. Atmanspacher and S. Maasen (Hoboken, NJ: John Wiley and Sons, Ltd), 341–362. doi: 10.1002/9781118865064.ch16

Anderson, M. L., and Oates, T. (2010). A critique of multi-voxel pattern analysis. *Proc. Annu. Meet. Cogn. Sci. Soc.* 32, 1511–1516.

Anderson, M. L., Kinnison, J., and Pessoa, L. (2013). Describing functional diversity of brain regions and brain networks. *NeuroImage* 73, 50–58. doi: 10.1016/ j.neuroimage.2013.01.071

Bair, W. (2005). Visual receptive field organization. *Curr. Opin. Neurobiol.* 15, 459–464. doi: 10.1016/j.conb.2005.07.006

Bargmann, C. I. (2012). Beyond the connectome: how neuromodulators shape neural circuits. *BioEssays* 34, 458–465. doi: 10.1002/bies.201100185

Barrett, L. (2011). *Beyond The Brain: How Body And Environment Shape Animal And Human Minds*. Princeton, NJ: Princeton University Press.

Bateson, P. P. G., and Gluckman, P. D. (2011). *Plasticity, Robustness, Development and Evolution*. Cambridge: Cambridge University Press.

Bechtel, W., and Abrahamsen, A. (2010). Dynamic mechanistic explanation: computational modeling of circadian rhythms as an exemplar for cognitive science. *Stud. Hist. Philos. Sci. Part A* 41, 321–333. doi: 10.1016/j.shpsa.2010.07.003

Bedny, M. (2017). Evidence from blindness for a cognitively pluripotent cortex. *Trends Cogn. Sci.* 21, 637–648. doi: 10.1016/j.tics.2017.06.003

Behrmann, M., and Plaut, D. C. (2014). Bilateral hemispheric processing of words and faces: evidence from word impairments in prosopagnosia and face impairments in pure alexia. *Cereb. Cortex (New York, N.Y.: 1991)* 24, 1102–1118. doi: 10.1093/cercor/bhs390

Bergeron, V. (2007). Anatomical and functional modularity in cognitive science: shifting the focus. *Philos. Psychol.* 20, 175–195. doi: 10.1080/09515080701197155

Berneiser, J., Jahn, G., Grothe, M., and Lotze, M. (2018). From visual to motor strategies: training in mental rotation of hands. *NeuroImage* 167, 247–255. doi: 10.1016/j.neuroimage.2016.06.014

Biduła, S. P., Przybylski, Ł, Pawlak, M. A., and Króliczak, G. (2017). Unique neural characteristics of atypical lateralization of language in healthy individuals. *Front. Neurosci.* 11:525. doi: 10.3389/fnins.2017.00525

Bolt, T., Anderson, M. L., and Uddin, L. Q. (2017). Beyond the evoked/intrinsic neural process dichotomy. *Netw. Neurosci.* 2, 1–22. doi: 10.1162/NETN_a_00028

Borra, E., and Luppino, G. (2018). Large-scale temporo–parieto–frontal networks for motor and cognitive motor functions in the primate brain. *Cortex* 118, 19–37. doi: 10.1016/j.cortex.2018.09.024

Bowren, M. D., Tranel, D., and Boes, A. D. (2021). Preserved cognition after right hemispherectomy. *Neurol. Clin. Pract.* 11, e906–e908.

Bracci, S., Daniels, N., and Op de Beeck, H. (2017). Task context overrules object- and category-related representational content in the human parietal cortex. *Cereb. Cortex* 27, 310–321. doi: 10.1093/cercor/bhw419

Broca, P. (1861). Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bull. Soc. Anat.* 6, 330–357.

Bruineberg, J., and Rietveld, E. (2019). What's inside your head once you've figured out what your head's inside of. *Ecol. Psychol.* 31, 198–217. doi: 10.1080/ 10407413.2019.1615204

Burnston, D. C. (2016b). A contextualist approach to functional localization in the brain. *Biol. Philos.* 31, 527–550. doi: 10.1007/s10539-016-9526-2

Burnston, D. C. (2016a). Computational neuroscience and localized neural function. *Synthese* 193, 3741–3762. doi: 10.1007/s11229-016-1099-8

Burnston, D. C. (2019). Getting over atomism: functional decomposition in complex neural systems. *Br. J. Philos. Sci.* 72, 743–772. doi: 10.1093/bjps/axz039

Burnston, D. C. (2021). Contents, vehicles, and complex data analysis in neuroscience. *Synthese* 199, 1617–1639. doi: 10.1016/j.neuroscience.2016.06.014

Buzsáki, G. (2020). The brain–cognitive behavior problem: a retrospective. *Eneuro* 7:ENEURO.0069-20.2020. doi: 10.1523/ENEURO.0069-20.2020

Buzsáki, G., and Tingley, D. (2018). Space and time: the hippocampus as a sequence generator. *Trends Cogn. Sci.* 22, 853–869.

Cabeza, R., and Nyberg, L. (2000). Imaging cognition II: an empirical review of 275 pet and fMRI studies. *J. Cogn. Neurosci.* 12, 1–47. doi: 10.1162/ 08989290051137585

Cartwright, N., Pemberton, J., and Wieten, S. (2020). Mechanisms, laws and explanation. *Eur. J. Philos. Sci.* 10, 1–19.

Chiel, H. J., and Beer, R. D. (1997). The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci.* 20, 553–557. doi: 10.1016/S0166-2236(97)01149-1

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philos. Trans. R. Soc. B* 362, 1585–1599. doi: 10.1098/rstb. 2007.2054

Clopath, C., Bonhoeffer, T., Hübener, M., and Rose, T. (2017). Variance and invariance of neuronal long-term representations. *Philos. Trans. R. Soc. B* 372:20160161. doi: 10.1098/rstb.2016.0161

Craver, C. F. (2014). "Levels," in *Open Mind*, eds T. Metzinger and J. M. Windt (Frankfurt am Main: MIND Group).

Craver, C. F., and Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *Br. J. Philos. Sci.* 71, 287–319.

Craver, C. F., and Tabery, J. (2015). *Mechanisms In Science*. Available online at: https://stanford.library.sydney.edu.au/archives/sum2017/entries/science-mechanisms/ (accessed May 28, 2020).

Çukur, T., Nishimoto, S., Huth, A. G., and Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* 16, 763–770. doi: 10.1038/nn.3381

De Brigard, F. (2017). Cognitive systems and the changing brain. *Philos. Explor.* 20, 224–241. doi: 10.1080/13869795.2017.1312503

de Wit, M. M., and Withagen, R. (2019). What should a "gibsonian neuroscience" look like? Introduction to the special issue. *Ecol. Psychol.* 31, 147–151. doi: 10.1080/10407413.2019.1615203

de Wit, M. M., de Vries, S., van der Kamp, J., and Withagen, R. (2017). Affordances and neuroscience: steps towards a successful marriage. *Neurosci. Biobehav. Rev.* 80, 622–629. doi: 10.1016/j.neubiorev.2017.07.008

de Wit, M. M., Van der Kamp, J., and Masters, R. S. W. (2012). Distinct task-independent visual thresholds for egocentric and allocentric information pick up. *Conscious. Cogn.* 21, 1410–1418. doi: 10.1016/j.concog.2012.07.008

Dehaene, S., and Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron* 56, 384–398. doi: 10.1016/j.neuron.2007.10.004

Dewhurst, J. (2018). Context-Sensitive ontologies for a non-reductionist cognitive neuroscience. *Aust. Philos. Rev.* 2, 224–228. doi: 10.1080/24740500.2018. 1552102

Dixon, R. A., Garrett, D. D., Bäckman, L., Stuss, D. T., and Winocur, G. (2008). Principles of compensation in cognitive neuroscience and neurorehabilitation. *Cogn. Neurorehabil.* 2, 22–38. doi: 10.1080/09602011.2014.1003947

Dotov, D. G. (2014). Putting reins on the brain. How the body and environment use it. *Front. Hum. Neurosci.* 8:795. doi: 10.3389/fnhum.2014.00795

Edelman, G. M., and Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13763–13768. doi: 10.1073/ pnas.231499798

Ekstrom, A. D., and Ranganath, C. (2018). Space, time, and episodic memory: the hippocampus is all over the cognitive map. *Hippocampus* 28, 680–687.

Eronen, M. I., and Bringmann, L. F. (2021). The theory crisis in psychology: how to move forward. *Perspect. Psychol. Sci.* 16, 779–788. doi: 10.1177/ 1745691620970586

Fancher, R. E. (1996). *Pioneers of Psychology*, 3rd Edn. London: Norton.

Fedorenko, E., and Blank, I. A. (2020). Broca's area is not a natural kind. *Trends Cogn. Sci.* 24, 270–284. doi: 10.1016/j.tics.2020.01.001

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philos. Sci.* 77, 419–456. doi: 10.1086/652964

Fotopoulou, A. (2014). Time to get rid of the 'modular' in neuropsychology: a unified theory of anosognosia as aberrant predictive coding. *J. Neuropsychol.* 8, 1–19. doi: 10.1111/jnp.12010

Friston, K. J., and Price, C. J. (2003). Degeneracy and redundancy in cognitive anatomy. *Trends Cogn. Sci.* 7, 151–152. doi: 10.1016/S1364-6613(03)00054-8

Gallivan, J. P., and Culham, J. C. (2015). Neural coding within human brain areas involved in actions. *Curr. Opin. Neurobiol.* 33, 141–149. doi: 10.1016/j.conb.2015.03.012

García, A. M., Sedeño, L., Herrera Murcia, E., Couto, B., and Ibáñez, A. (2017). A lesion-proof brain? Multidimensional sensorimotor, cognitive, and socio-affective preservation despite extensive damage in a stroke patient. *Front. Aging Neurosci.* 8:335. doi: 10.3389/fnagi.2016.00335

Gardner, M. R., Brazier, M., Edmonds, C. J., and Gronholm, P. C. (2013). Strategy modulates spatial perspective-taking: evidence for dissociable disembodied and embodied routes. *Front. Hum. Neurosci.* 7:457. doi: 10.3389/fnhum.2013.00457

Garson, J. (2016). *A Critical Overview of Biological Functions*. Cham: Springer International Publishing, doi: 10.1007/978-3-319-32020-5

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (2019). *Cognitive Neuroscience: The Biology Of The Mind*, 5th Edn. New York, NY: W.W. Norton and Company.

Genon, S., Reid, A., Langner, R., Amunts, K., and Eickhoff, S. B. (2018). How to characterize the function of a brain region. *Trends Cogn. Sci.* 22, 350–364. doi: 10.1016/j.tics.2018.01.010

Gibson, J. J. (1966). *The Senses Considered As Perceptual Systems*. Boston, MA: Houghton Mifflin.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171–178. doi: 10.1038/nature18933

Goldenberg, G., and Randerath, J. (2015). Shared neural substrates of apraxia and aphasia. *Neuropsychologia* 75, 40–49. doi: 10.1016/j.neuropsychologia.2015.05.017

Harel, A., Kravitz, D. J. D., and Baker, C. I. C. (2014). Task context impacts visual object processing differentially across the cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, E962–E971. doi: 10.1073/pnas.1312567111

Harnish, S., Meinzer, M., Trinastic, J., Fitzgerald, D., and Page, S. (2014). Language changes coincide with motor and fMRI changes following upper extremity motor therapy for hemiparesis: a brief report. *Brain Imaging Behav.* 8, 370–377. doi: 10.1007/s11682-011-9139-y

Hartwigsen, G. (2018). Flexible redistribution in cognitive networks. *Trends Cogn. Sci.* 22, 687–698. doi: 10.1016/j.tics.2018.05.008

Hartwigsen, G., and Saur, D. (2019). Neuroimaging of stroke recovery from aphasia – Insights into plasticity of the human language network. *NeuroImage* 190, 14–31. doi: 10.1016/j.neuroimage.2017.11.056

Hasson, U., Nastase, S. A., and Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* 105, 416–434. doi: 10.1016/j.neuron.2019.12.002

Humphreys, G. F., Ralph, M. A. L., and Simons, J. S. (2021). A unifying account of angular gyrus contributions to episodic and semantic cognition. *Trends Neurosci.* 44, 452–463. doi: 10.1016/j.tins.2021.01.006

Hutto, D. D., Peeters, A., and Segundo-Ortin, M. (2017). Cognitive ontology in flux: the possibility of protean brains. *Philos. Explor.* 9795, 1–17. doi: 10.1080/13869795.2017.1312502

Jeffery, K. J. (2018). The hippocampus: from memory, to map, to memory map. *Trends Neurosci.* 41, 64–66.

Jiang, H., Stanford, T. R., Rowland, B. A., and Stein, B. E. (2021). Association cortex is essential to reverse hemianopia by multisensory training. *Cereb. Cortex* 31, 5015–5023. doi: 10.1093/cercor/bhab138

Jones, M. (2018). Numerals and neural reuse. *Synthese* 197, 3657–3681. doi: 10.1007/s11229-018-01922-y

Journal of Cognitive Neuroscience (2021). Available online at: https://direct.mit.edu/jocn [Accessed May 18, 2021].

Kaiser, M. I. (2018). "The components and boundaries of mechanisms," in *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, eds S. Glennan and P. Illari (New York, NY: Routledge).

Kästner, L. (2017). *Philosophy of Cognitive Neuroscience: Causal Explanations, Mechanisms and Experimental Manipulations*. Berlin: De Gruyter.

Khalidi, M. A. (2017). Crosscutting psycho-neural taxonomies: the case of episodic memory. *Philos. Explor.* 20, 191–208. doi: 10.1080/13869795.2017.1312501

King, M., Hernandez-Castillo, C. R., Poldrack, R. A., Ivry, R. B., and Diedrichsen, J. (2019). Functional boundaries in the human cerebellum revealed by a multi-domain task battery. *Nat. Neurosci.* 22, 1371–1378. doi: 10.1038/s41593-019-0436-x

Kiverstein, J. D., and Miller, M. (2015). The embodied brain: towards a radical embodied cognitive neuroscience. *Front. Hum. Neurosci.* 9:237. doi: 10.3389/fnhum.2015.00237

Klein, C. (2012). Cognitive ontology and region- versus network-oriented analyses. *Philos. Sci.* 79, 952–960. doi: 10.1086/667843

Knops, A., Thirion, B., Hubbard, E. M., Michel, V., and Dehaene, S. (2009). Recruitment of an area involved in eye movements during mental arithmetic. *Science* 324, 1583–1585. doi: 10.1126/science.1171599

Kolb, B., and Gibb, R. (2013). Searching for the principles of brain plasticity and behavior. *Cortex* 58, 251–260.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., Maciver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490. doi: 10.1016/j.neuron.2016.12.041

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., and Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* 17, 26–49. doi: 10.1016/j.tics.2012.10.011

Kriegeskorte, N., and Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annu. Rev. Neurosci.* 42, 407–432.

Kriegeskorte, N., and Wei, X. X. (2021). Neural tuning and representational geometry. *arXiv* [Preprint]. arXiv: 2104.09743.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Lee, J., and Dewhurst, J. (2021). The mechanistic stance. *Eur. J. Philos. Sci.* 11, 1–21.

Maguire, E. A., Intraub, H., and Mullally, S. L. (2016). Scenes, spaces, and memory traces: what does the hippocampus do? *Neuroscientist* 22, 432–439.

Marder, E., Goeritz, M. L., and Otopalik, A. G. (2015). Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. *Curr. Opin. Neurobiol.* 31, 156–163. doi: 10.1016/j.conb.2014.10.012

Marr, D. (1982/2010). *Vision: A Computational Investigation Into The Human Representation And Processing Of Visual Information*. Cambridge, MA: MIT Press.

Matheson, H. E., Garcea, F. E., and Buxbaum, L. J. (2021). Scene context shapes category representational geometry during processing of tools. *Cortex* 141, 1–15. doi: 10.1016/j.cortex.2021.03.021

Maurer, A. P., and Nadel, L. (2021). The continuity of context: a role for the hippocampus. *Trends Cogn. Sci.* 25, 187–199.

McIntosh, A. R. (1999). Mapping cognition to the brain through neural interactions. *Memory* 7, 523–548. doi: 10.1080/096582199387733

McIntosh, A. R. (2000). Towards a network theory of cognition. *Neural Netw.* 13, 861–870. doi: 10.1016/S0893-6080(00)00059-9

McIntosh, A. R. (2004). Contexts and catalysts: a resolution of the localization and integration of function in the brain. *Neuroinformatics* 2, 175–182. doi: 10.1385/NI:2:2:175

Menary, R. (2015). "Mathematical cognition—a case of enculturation," in *Open MIND*, eds T. Metzinger and J. M. Windt (Frankfurt am Main: MIND Group), 25. doi: 10.15502/9783958570818

Merabet, L. B., Hamilton, R., Schlaug, G., Swisher, J. D., Kiriakopoulos, E. T., Pitskel, N. B., et al. (2008). Rapid and reversible recruitment of early visual cortex for touch. *PLoS One* 3:e3046. doi: 10.1371/journal.pone.0003046

Mesulam, M.-M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Ann. Neurol.* 28, 597–613. doi: 10.1002/ana.410280502

Miłkowski, M. (2013). *Explaining The Computational Mind*. Cambridge, MA: The MIT Press.

Miller, E. K. (2000). The prefontral cortex and cognitive control. *Nat. Rev. Neurosci.* 1, 59–65. doi: 10.1038/35036228

Mogensen, J. (2011). Reorganization of the injured brain: implications for studies of the neural substrate of cognition. *Front. Psychol.* 2:7. doi: 10.3389/fpsyg.2011.00007

Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., and Matusz, P. J. (2015). The multisensory function of the human primary visual cortex. *Neuropsychologia* 1, 1–9. doi: 10.1016/j.neuropsychologia.2015.08.011

Noppeney, U., Friston, K. J., and Price, C. J. (2004). Degenerate neuronal systems sustaining cognitive functions. *J. Anat.* 205, 433–442. doi: 10.1111/j.0021-8782.2004.00343.x

Noppeney, U., Penny, W. D., Price, C. J., Flandin, G., and Friston, K. J. (2006). Identification of degenerate neuronal systems based on intersubject variability. *NeuroImage* 30, 885–890. doi: 10.1016/j.neuroimage.2005.10.010

Parkinson, C., Liu, S., and Wheatley, T. (2014). A common cortical metric for spatial, temporal, and social distance. *J. Neurosci.* 34, 1979–1987. doi: 10.1523/JNEUROSCI.2159-13.2014

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nat. Rev. Neurosci.* 9, 148–158.

Piccinini, G. (2022). Situated neural representations: solving the problems of content. *Front. Neurorobotics* 16:846979. doi: 10.3389/fnbot.2022.846979

Piccinini, G., and Craver, C. F. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183, 283–311. doi: 10.1007/s11229-011-9898-4

Piccinini, G., and Scarantino, A. (2011). Information processing, computation, and cognition. *J. Biol. Phys.* 37, 1–38. doi: 10.1007/s10867-010-9195-3

Piccinini, G., and Shagrir, O. (2014). Foundations of computational neuroscience. *Curr. Opin. Neurobiol.* 25, 25–30. doi: 10.1016/j.conb.2013.10.005

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10, 59–63. doi: 10.1016/j.tics.2005.12.004

Poldrack, R. A. (2010). Mapping mental function to brain structure: how can cognitive neuroimaging succeed? *Perspect. Psychol. Sci.* 5, 753–761. doi: 10.1177/1745691610388777

Poldrack, R. A., and Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annu. Rev. Psychol.* 67, 1–26. doi: 10.1146/annurev-psych-122414-033729

Price, C. J. (2018). The evolution of cognitive models: from neuropsychology to neuroimaging and back. *Cortex* 107, 37–49. doi: 10.1016/j.cortex.2017.12.020

Price, C. J., and Friston, K. J. (2002). Degeneracy and cognitive anatomy. *Trends Cogn. Sci.* 6, 416–421. doi: 10.1016/S1364-6613(02)01976-9

Price, C. J., and Friston, K. J. (2005). Functional ontologies for cognition: the systematic definition of structure and function. *Cogn. Neuropsychol.* 22, 262–275. doi: 10.1080/02643290442000095

Price, C. J., Hope, T. T. M., and Seghier, M. L. (2016). Ten problems and solutions when predicting individual outcome from lesion site after stroke. *NeuroImage* 145, 200–208. doi: 10.1016/j.neuroimage.2016.08.006

Primaßin, A., Scholtes, N., Heim, S., Huber, W., Binkofski, F., and Werner, C. J. (2015). Determinants of concurrent motor and language recovery during intensive therapy in chronic stroke patients: four single-case studies. *Front. Neurol.* 6:215. doi: 10.3389/fneur.2015.00215

Raja, V. (2021). Resonance and radical embodiment. *Synthese* 199, 113–141. doi: 10.1007/s11229-020-02610-6

Raja, V., and Anderson, M. L. (2021). "Behavior considered as an enabling constraint," in *Neural Mechanisms*, Vol. 17, eds F. Calzavarini and M. Viola (Cham: Springer International Publishing), 209–232. doi: 10.1007/978-3-030-54092-0_10

Raymer, A. M. (1999). Crossed apraxia: implications for handedness. *Cortex* 35, 183–199. doi: 10.1016/S0010-9452(08)70793-7

Renoult, L., Irish, M., Moscovitch, M., and Rugg, M. D. (2019). From knowing to remembering: the semantic–episodic distinction. *Trends Cogn. Sci.* 23, 1041–1057.

Rule, M. E., O'Leary, T., and Harvey, C. D. (2019). Causes and consequences of representational drift. *Curr. Opin. Neurobiol.* 58, 141–147.

Ryan, K. J. J., and Gallagher, S. (2020). Between ecological psychology and enactivism: is there resonance? *Front. Psychol.* 11:1147. doi: 10.3389/fpsyg.2020.01147

Salehi, M., Greene, A. S., Karbasi, A., Shen, X., Scheinost, D., and Constable, R. T. (2020). There is no single functional atlas even for a single individual: functional parcel definitions change with task. *NeuroImage* 208:116366. doi: 10.1016/j.neuroimage.2019.116366

Seifert, L., Komar, J., Araújo, D., and Davids, K. (2016). Neurobiological degeneracy: a key property for functional adaptations of perception and action to constraints. *Neurosci. Biobehav. Rev.* 69, 159–165. doi: 10.1016/j.neubiorev.2016.08.006

Shea, N. (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.

Shine, J. M., Eisenberg, I., and Poldrack, R. A. (2016). Computational specificity in the human brain. *Behav. Brain Sci.* 39:e131. doi: 10.1017/S0140525X1500165X

Sporns, O. (2011). *Networks Of The Brain*. Cambridge, MA: The MIT Press.

Sporns, O. (2012). *Discovering The Human Connectome*. Cambridge, MA: MIT Press.

Stanley, M. L., Gessell, B., and De Brigard, F. (2019). Network modularity as a foundation for neural reuse. *Philos. Sci.* 86, 23–46. doi: 10.1086/701037

Stanley, M. L., Moussa, M. N., Paolini, B., Lyday, R. G., Burdette, J. H., and Laurienti, P. J. (2013). Defining nodes in complex brain networks. *Front. Comput. Neurosci.* 7:169. doi: 10.3389/fncom.2013.00169

Stoll, H., de Wit, M. M., Middleton, E. L., and Buxbaum, L. J. (2021). Treating limb apraxia *via* action semantics: a preliminary study. *Neuropsychol. Rehabil.* 31, 1145–1162. doi: 10.1080/09602011.2020.1762672

Tang, Y., Zhang, W., Chen, K., Feng, S., Ji, Y., Shen, J., et al. (2006). Arithmetic processing in the brain shaped by cultures. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10775–10780. doi: 10.1073/pnas.0604416103

Uddin, L. Q., Kinnison, J., Pessoa, L., and Anderson, M. L. (2014). Beyond the tripartite cognition—emotion—interoception model of the human insular cortex. *J. Cogn. Neurosci.* 26, 16–27. doi: 10.1162/jocn_a_00462

Uddin, L. Q., Yeo, B. T. T., and Spreng, R. N. (2019). Towards a universal taxonomy of macro-scale functional human brain networks. *Brain Topogr.* 32, 926–942. doi: 10.1007/s10548-019-00744-6

Uttal, W. R. (2001). *The New Phrenology: The Limits Of Localizing Cognitive Processes In The Brain*. Cambridge, MA: MIT Press.

van der Weel, F. R., Agyei, S. B., and van der Meer, A. L. H. (2019). Infants' brain responses to looming danger: degeneracy of neural connectivity patterns. *Ecol. Psychol.* 31, 182–197. doi: 10.1080/10407413.2019.1615210

van Orden, G., Hollis, G., and Wallot, S. (2012). The blue-collar brain. *Front. Physiol.* 3:207. doi: 10.3389/fphys.2012.00207

Vingerhoets, G., Alderweireldt, A. S., Vandemaele, P., Cai, Q., Van der Haegen, L., Brysbaert, M., et al. (2013). Praxis and language are linked: evidence from co-lateralization in individuals with atypical language dominance. *Cortex* 49, 172–183. doi: 10.1016/j.cortex.2011.11.003

Viola, M. (2020). Beyond the platonic brain: facing the challenge of individual differences in function-structure mapping. *Synthese* 199, 2129–2155. doi: 10.1007/s11229-020-02875-x

Viola, M., and Zanin, E. (2017). The standard ontological framework of cognitive neuroscience: some lessons from broca's area. *Philos. Psychol.* 5089, 1–25. doi: 10.1080/09515089.2017.1322193

Wernicke, C. (1874/1969). "The symptom complex of aphasia," in *Proceedings of the Boston Colloquium for the Philosophy of Science 1966/1968*, eds R. S. Cohen and M. W. Wartofsky (Dordrecht: Springer Netherlands), 34–97. doi: 10.1007/978-94-010-3378-7_2

Yassa, M. A., and Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends Neurosci.* 34, 515–525.

Zednik, C. (2018). "Mechanisms in cognitive science," in *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, eds P. M. Illari and S. Glennan (London: Routledge), 389–400.

Zednik, C. (2019). "Computational cognitive neuroscience," in *The Routledge Handbook of the Computational Mind*, eds M. Sprevak and M. Colombo (Abingdon: Routledge).

Zerilli, J. (2019). Neural reuse and the modularity of mind: where to next for modularity? *Biol. Theory* 14, 1–20. doi: 10.1007/s13752-018-0309-7

Zerilli, J. (2020). *The Adaptable Mind: What Neuroplasticity and Neural Reuse Tells Us About Language and Cognition*. Oxford: Oxford University Press.

frontiers | Frontiers in Psychology

TYPE Review
PUBLISHED 13 September 2022
DOI 10.3389/fpsyg.2022.823971

A glimpse into social perception in light of vitality forms

Qingming Liu[1,2], Jinxin Zhang[3], Da Dong[1,2*†] and Wei Chen[1,2,4*†]

[1]Center for Brain, Mind and Education, Shaoxing University, Shaoxing, China, [2]Department of Psychology, Shaoxing University, Shaoxing, China, [3]Department of Psychology, Zhejiang Normal University, Jinhua, China, [4]Interdisciplinary Center for Philosophy and Cognitive Sciences, Renmin University of China, Beijing, China

Check for updates

OPEN ACCESS

EDITED BY
Guy Dove,
University of Louisville, United States

REVIEWED BY
Giuseppe Di Cesare,
Italian Institute of Technology (IIT), Italy
Eros Moreira De Carvalho,
Federal University of Rio Grande do Sul, Brazil
Jean-Marie C. Chevalier,
Université Paris-Est Créteil Val de Marne, France

*CORRESPONDENCE
Da Dong
feixiandd@163.com
Wei Chen
anti-monist@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Theoretical and Philosophical Psychology,
a section of the journal
Frontiers in Psychology

RECEIVED 28 November 2021
ACCEPTED 15 August 2022
PUBLISHED 13 September 2022

CITATION
Liu Q, Zhang J, Dong D and Chen W (2022) A glimpse into social perception in light of vitality forms.
Front. Psychol. 13:823971.
doi: 10.3389/fpsyg.2022.823971

COPYRIGHT
© 2022 Liu, Zhang, Dong and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The American psychoanalyst and developmental psychologist Daniel Stern's idea of vitality forms might suggest a new solution to explain how other minds are intensely expressed in their actions. Vitality forms characterize the expressive style of actions. The effective perception of vitality forms allows people to recognize the affective states and intentions of others in their actions, and could even open the possibility of properties of objects that are indicated by the given actions. Currently, neurophysiological studies present that there might be a neural mirror mechanism in the dorso-central insula (DCI), middle cingulate cortex (MCC), and other related cerebral areas, which serve to preferably perceive and deliver vitality forms of actions. In this article, possible types of vitality forms related to other minds, which have been brought to particular attention in recent years, have been collected and discussed in the following four areas: (1) Vitality forms on understanding non-verbal intention, (2) on understanding verbal intention, (3) vitality forms as grounding social cognition, and (4) as grounding social emotion. These four areas, however, might refer to an entirety of a binary actor-observer communicative landscape. In this review, we try to simplify the analysis by relying on two fundamental dimensions of criteria: first, the idea of vitality forms is conceived as the most basic way of observing subsequent higher-order dimensions of action, that is, understanding intention in the style of action. Thus, in the first two subsections, the relationships between vitality forms and their roles in understanding non-verbal and verbal intention have been discussed. Second, vitality forms could also be conceived as background conditions of all the other mental categories, that is, vitality forms can ground cognition and emotion in a social context. In the second dimension, the existence of social cognition or emotion depends on the existence of the stylistic kinematics of action. A grounding relation is used to distinguish a ground, that is, vitality forms, and its grounded mental categories. As relating with the domain of social perception, in this review, it has been discussed vitality forms possibly could ground social cognition and social emotion, respectively.

KEYWORDS

vitality forms, social perception, intention understanding, social affordances, social emotion

Introduction

Epigraph

> The expressiveness of vitality affects can be likened to that of a puppet show. The puppets have little or no capacity to express categories of affects by way of facial signals, and their repertoire of conventionalized gestural or postural affect signals is usually impoverished. It is from the way they move in general that we infer the different vitality affects from the activation contours they trace. Most often, the characters of different puppets are largely defined in terms of particular vitality affects: one may be lethargic, with drooping limbs and hanging head, another forceful, and still another jaunty.
>
> (Stern, 1998, p. 56)

Introduction

Adults' affective experiences have been categorized to a large extent. These are, in a certain sense, a result of social life over a period of time. In the categories of affects, happiness and sadness (the two extremes of affective states) would not be confused for a disciplined adult. It seems that each category owns an unambiguous definition and corresponds to a distinct experiential state. When an adult cries, she cries for a purpose, so to speak; more importantly, the purpose can be articulated in a verbal way. When an adult acts, she acts intentionally toward an object, no matter actual or virtual. Then, consider the affective lives of young infants. In contrast, categories of affects in infants are difficult to be recognized clearly and distinctly by an arbitrary external observer. Babies cry loudly and non-verbally; babies act intensely, and in most instances, use their entire body. It seems that there is an explanatory gap between adults' categories of affects and infants' somehow undifferentiated primordial feelings along the lines of orthodox affective category theory; representatives of the latter include Charles Darwin (1872).

The American psychoanalyst and developmental psychologist Daniel Stern in a radically different sense disagreed with affective category theory. Stern once summarized the latter's view,

> "...discrete categories of affect...each of these had an innate discrete facial display and a distinct quality of feeling and that these innate patterns evolved as social signals "understood" by all members to enhance species survival" (Stern, 1998, p. 54–55).

On the contrary, Stern founded affective *gestalt* (form) theory, or what he precisely named "vitality affect" and then the final generalized theory, namely "vitality form." In this transition of concepts (affect → form), Stern wanted to go beyond the mere affective spheres of one category of the human mind. The concept of vitality forms was then conceived by Stern as background conditions of all mental categories (including attention, emotion, volition, etc.). Practically, this idea is based on Stern's long-term close observation of young infants' interpersonal world, especially parental–infant dyadic interactions, for example, the affective bonding relations of infants with their mothers (Stern, 1977). A common definition of vitality forms goes like this: this crucial term is directly related to basic kinematic characteristics of actions, which represent lively experiential states of the mind/the agent. Specifically, it refers to how the action is carried out in space and time, that is, which refers to the "how" dimension of the action. Each vitality form has a specific kinematic "contour" which can be detected according to the kinematics of actions. Depending on people's affective or cognitive states, their actions may take on different styles or manners of vitality forms. For example, the action of gripping a cup can be "strong" or "delicate"; the action of touching a cat can be "rude" or "gentle" (Stern, 2010).

Presently, there is an increasingly influential current in social perception debates that seeks inspiration from Stern's vitality forms (cf. Di Cesare et al., 2014; Marraffa and Meini, 2019; Casartelli et al., 2020a). The idea of vitality forms also attracts mirror neurons' discoverer, the neuroscientist Giacomo Rizzolatti's attention. Through the efforts of Rizzolatti, Di Cesare, and their research team, the issue of neural bases of vitality forms, relating to the neural mirroring mechanism of social perception, is also discussed in the realm at present (cf. Gallese and Rochat, 2018; Di Cesare et al., 2020; Rizzolatti et al., 2021).

Generally, there are three traditional theoretical approaches to the current debates on the philosophy of social perception: Rationality theory, theory-theory, and simulation theory (Apperly, 2010). Rationality theory supposes that each human person is a rational being; reading other minds is a process of actively rationalizing others' beliefs (Goldman, 2006). Theory-theory argues that people explain and forecast the behavior of others by applying a set of folk psychological theories (Davies and Stone, 1995); thus, the key point of this approach relates to the internal construction of a *theory* of one's own. Simulation theory assumes that people's own behaviors and psychological relationships are similar to those of others. Thus, in order to explain the behavior of others, in a certain way, people *simulate* the mental states of others in their own minds (cf. Hurley, 2008). Currently, orthodox views of other minds as indirectly inferred activities have been challenged by direct social perception theory already; that is, it is possible to directly *see* the mental states of other people (cf. Gallagher, 2020; Krueger, 2021). Besides, in recent times, experimental psychologists, such as Cristina Becchio, have already tried to design a series of down-to-earth experiments to defend this new approach (Becchio et al., 2018).

This article proposes that the idea of vitality forms allows people to *see* others' minds in their actions directly. In other words, it is possible to understand the attitudes or affective states of others by simply observing their actions (e.g., by listening to their varying tone of voice).[1] Vitality forms represent the way an action could be performed. Imagine an agent interacts with the other agent in a dyad. According to their attitude (positive or negative), interoceptive state (feel good or feel bad), and emotion (happy or sad), the final action is that they will pass an object gently or rudely. Note that during this *very* action, so to speak, there is always a one-to-one correspondence between the intention of the agent (for passing an object) and the how-dimension of action (vitality forms, i.e., how an agent passes this object). Gallagher (2020) claims that the concept of vitality forms possibly is the prerequisite of subsequent contents and intentions of social perception; he highlights that only in this context people can talk about "an emergence of meaning, a meaning that emerges in the interaction itself" (Gallagher, 2020, p. 161; see also Gallagher et al., 2022). Krueger (2021) tries to defend direct social perception with the aid of vitality forms. He argues that people see others' actions "performed *in a particular sort of way*, as embodying a particular manner or style" (Krueger, 2021, p. S373). That is, people see vitality forms directly, observing the mind directly in action (Krueger, 2021, p. S368). Combining those converging reflections of Gallagher, Krueger, and others, the approach of action-how (vitality forms) is radically different from action-what (contents) and action-why (intentions). In a certain sense, it defends the direct social perception approach (as well as challenges the orthodox mindreading theory). In the following paragraphs, we have collected and reviewed the existing relevant empirical research with a synopsis to clarify how the mental states of others are directly represented in the perception of others' expressive kinematics through vitality forms of action. The scope of this paper is to reveal the significance of vitality forms for kinematic interpersonal relations.

Burgeoning areas of vitality forms to social perception

In Stern's approach, social perception activities take parental–infant dyadic relationships as proto-types. Vitality forms occupy a new dimension apart from affective category theory, actions *per se* convey enough in revealing the kinematic signatures of social perception. To adults, previous parental–infant intersubjective activities may remain components in the adult stage. Vitality forms should be seen as more fundamental than derivative categories of affects, and contents or intentions of mindreading. Thus, more or less metaphorically, not only infants but also adults are immersed in "feelings of vitality":

> "Different feelings of vitality can be expressed in a multitude of parental acts that do not qualify as "regular" affective acts: how the mother picks up baby, folds the diapers, grooms her hair or the baby's hair, reaches for a bottle, unbuttons her blouse. The infant is immersed in these "feelings of vitality"" (Stern, 1998, p. 54).

Another important person in the realm of developmental intersubjectivity studies, Trevarthen (2013) applauds highly of Stern's last monograph on forms of vitality, "has made a bible for all humanistic and art therapies" (Trevarthen, 2013, p. 44). In a sense, vitality forms play a constitutive role in the bottom level of the stratified structures of intersubjectivity. Combining Trevarthen's concept of primary intersubjectivity (the stage of young infants before 9 months), vitality forms are the first perceivable signatures that young infants can *see* (directly) (cf. Gallagher, 2020). However, it is far from clear "how body, brain, and mind collaborate in its shared vitality" (Trevarthen, 2019, p. 31) based on the current empirical studies of vitality forms.

Further, in this article, possible types of vitality forms related to other minds are collected and discussed in the following four areas across two dimensions of criteria, which have been brought to particular attention in recent years: (1) Vitality forms on understanding non-verbal intention, (2) on understanding verbal intention, (3) vitality forms as grounding social cognition, and (4) as grounding social emotion. These four areas might refer to an entirety of a binary actor–observer communicative landscape. Consider that a person performing a certain action is being observed, or, in other words, there exists an ongoing non-verbal/verbal communication between two actors (also as observers). There are five factors that can possibly be considered within this landscape: factors involving the *actor*, factors involving the *observer*, factors involving the *action*, factors involving the *relationship* between the actor and the observer, and factors involving the *context* (Kemmerer, 2021). All intersubjective characteristics of ongoing actions within a dyad can be seen as instantiations and their combinations of these five factors.

Thus, we try to simplify the analysis by relying on two fundamental dimensions: first, the idea of vitality forms is conceived as the most basic way of *seeing* subsequent higher-order dimensions of action, that is, understanding intention in the *style* of action (Stern, 2010; Krueger, 2021). Thus, the following first two subsections discuss the relationships between vitality forms and their role in understanding non-verbal and verbal intentions.

Second, vitality forms could also be conceived as background conditions of all the other mental categories, that is, vitality forms can *ground* cognition and emotion in

[1] We thank a referee for raising this issue.

social context.[2] In the second dimension, the existence of social cognition or emotion depends on the existence of the stylistic kinematics of action. A grounding relation is used to distinguish a ground, namely vitality forms, and its grounded mental categories (cf. Barsalou, 2010). In relating to the domain of social perception, we discuss vitality forms that possibly could ground social cognition and social emotion, respectively[3] (see **Table 1**).

The relationships between vitality forms and these four burgeoning areas are briefly reviewed in this article. Increasing evidence suggests that one type of non-inferentialist social perception theory allied with vitality forms promises to support a direct social perception approach. However, this enterprise underlying inferentialist–non-inferentialist social perception debate is still forthcoming because different studies lack essential theoretical relevance until the present time. In a long-range sense, the aim of the article is to make headway on the possible cooperation of vitality forms and direct social perception, that is, seeing the mind in the vitality forms of action directly.

Vitality forms on understanding non-verbal intention

When an agent has a positive or negative attitude (in other words, affective valence) toward the other agent, they will perform the following actions in a gentle or rude way, regardless of the type of action contents or intentions (i.e., taking a bottle to drink for themselves, or to throw it in the face of others). Each action is characterized by three different aspects: the content (action-what: Taking the bottle), the intention (action-why: Taking the bottle because of thirst), and the vitality forms (action-how: Taking the bottle hastily rather than tardily) (Stern, 2010).[4] For example, observing a person who greets you in a distance, in a certain sense *via* their action kinematics can help understand if they feel good or not, or if they are happy or sad. The same thing happens when answering the phone. It is possible to comprehend how the other person feels, directly, by hearing the tone contours of their voice regardless of verbal information. In one study specifically related to auditory vitality forms, Di Cesare et al. (2022) imply that communicative intention is a necessary condition for processing one of the candidates of neural bases of vitality forms, that is, dorso-central insula (DCI).

TABLE 1 Four burgeoning areas that vitality forms serve as understanding and grounding, respectively.

Vitality forms		
Understanding	Non-verbal intention	Verbal intention
Grounding	Social cognition	Social emotion

Prior to introducing the term "vitality forms," several empirical studies in cognitive psychology may somehow confirm the significance of kinematics of behavior in effective communicative interactions. Becchio et al. (2008) show that the kinematics of the movements is sensitive to social intention, and the kinematic patterns of the movements performed under two conditions (social condition and single-agent condition) are significantly different. Sartori et al. (2011) imply that action observers can distinguish between cooperative, competitive, and personal actions simply by observing the initial stretch of the action to the grasp stage; besides, an intention identification study comparing video and point light source clips by Manera et al. (2010) seem to also confirm this conclusion (also see Cavallo et al., 2016; Koul et al., 2016). Cavina-Pratesi et al. (2011) find that skilled magicians are sensitive to kinematic differences between grasping movements of real and imaginary objects. Ansuini et al. (2016) suggest that in a certain sense, the kinematic characteristics of pantomime movements may reflect the weight information of the target object. Podda et al.'s (2017) result shows that the action observer could tell whether a non-existent object is light or heavy on the basis of the kinematics of the observed action. However, it should be noted that all the examples provided in this passage do not assume a direct link between those previous kinematics research studies and vitality forms.

The relationship between autism spectrum disorder (ASD) and perceiving vitality forms has sought attention from an increasing number of researchers (Bystrom et al., 2019). ASD is a neurodevelopmental disorder with various clinical features, including deficits in social skills, verbal and non-verbal communication, and restricted and repetitive behaviors (cf. American Psychiatric Association, 2013, p. 40). There exists evidence that individuals with ASD are poor at recognizing emotional expressions in others (Atkinson, 2009). This may be linked to their difficulty in recognizing the vitality forms of other people's movements. Rochat et al. (2013) then investigate this hypothesis by asking children with ASD and healthy children to judge whether the two observed movements are similar or different in terms of vitality forms. The results show a distinct separation between the two tasks; autistic children have a remarkable impairment in recognizing the vitality forms of other people's actions. Casartelli et al. (2020a) compare autistic children with healthy children in different vitality forms (mild or rude) when performing hand movements (for example, placing a bottle, throwing a ball, or giving cookies). The results present those children with ASD show significant difficulty in changing

[2] We agree with Kemmerer that perceiving vitality forms is also modulated by "the situational context in which actions take place" (Kemmerer, 2021, p. 15).

[3] For other usages of grounding relation in cognitive science, for example, Larsson (2018) distinguishes two interrelated senses of grounding: communicative grounding, primary pragmatic actions (like utterances) in dialogue ground all the other activities of verbal communications; symbol grounding, symbolized language (like words) is essentially grounded in the world.

[4] We thank a referee for raising this issue.

their vitality forms while maintaining the same type of action. That is, they cannot express the vitality forms as a component of action, which is different from the goal of action. Effective social interaction is not only about understanding others, but also about making others understand us. It is a bilateral interaction process. Casartelli et al. (2020b) then further investigate whether healthy children could recognize vitality forms in autistic children. Based on the previous study, they find that healthy children have significant difficulties in understanding vitality forms displayed by autistic children, and yet still perform poorly with the aid of information feedback (for the latest review of ASD studies in the light of vitality forms, see Rochat and Gallese, 2022).

Vitality forms on understanding verbal intention

In our view, the role of vitality forms in understanding non-verbal intention is of primary significance here. Young infants have poor verbal communicative capacities. The vast majority of the communicative activities within a parental–infant dyad are presented in a non-verbal way. Besides, in activities of verbal communication, adult people to a large degree can ignore those kinematic clues considered here that have no direct relation with semantic information of human speech (such as the pitch, accent, and tone of voice). So, we turn to focus on young infants' vitality forms of speech, since they depend largely on the how-dimension of action for daily communication and developing their social cognition capacities. Generally, in relation to vitality forms in the realm of social perception, the studies of non-verbal intention and those of verbal intention are somehow separated.

In the process of human communication, verbal interaction and verbal intention understanding represent the highest stage of communication. Vitality forms could be transmitted not only through observable gestures and actions, but also through verbal signs. According to the speaker's attitude toward the listener, the listener can perceive the speaker mildly or rudely. Therefore, words that convey the form of vitality could allow the speaker to convey their inner mental state, and meanwhile, also allow the listener to understand the speaker's affective feelings. Stern once asserted that infants could perceive the vitality forms of speech. For example, in mothers' interactions with children, they often use a typical infant language for pronunciation. Specifically, the mothers can slow down the pronunciation of words to adapt their language to the children's perceiving limits (Stern, 1998).

One article uses functional magnetic resonance imaging (fMRI) technology to study how people can recognize the inner state of others by listening to their speech (Di Cesare et al., 2016). Participants are asked to listen to action verbs in three different conditions: a human voice delivers the verb in a rude or gentle manner, a robot voice delivers the same verb but without vitality forms, and a scrambled version of the same verb is delivered by a human voice. Consistent with previous studies on encoding vitality forms, this study finds specific activation of the central part of the insula when listening to human voices which possibly convey specific vitality forms. Both posterior parts of the left inferior frontal gyrus and anterior parietal motion circuits are activated when hearing human and robotic voices, which are typically activated when observing and performing arm movements. In all those three cases, the superior temporal gyrus is activated on both sides. The conclusion of the study is that the central part of the insula is a key area for processing vitality forms, and it is capable of understanding vitality forms regardless of the way they are communicated or expressed. In a subsequent fMRI study, the same research team tries to determine that the DCI, which is involved in the verbal perception of vitality forms, also becomes active during the imaginary process of generating action verbs of different vitality forms (Di Cesare et al., 2017a). The experiment is based on the fMRI technique. Due to technical reasons, movement cannot be studied. So, researchers cannot directly study the form of vitality from the form of speech. The researchers use the motor imagination of the same motion verbs as a strategy to evaluate the possible activation of the insular cortex in the process of generating vitality forms. In the speech condition, the participants are asked to listen to or imagine themselves speaking softly or rudely. In the action condition, the subjects are asked to observe or imagine whether their actions were mild or rude. The results show that compared to the control condition, the DCI is activated in both the speech and action tasks. Indeed, it has also been shown that the circuits activated by motor imagination are the same as those activated during action execution, except for the primary motor cortex (Jeannerod, 1995).

Vitality forms as grounding social cognition

Behavioral research has proved that even if there is no actual intention to act, the observation and indication of objects can even trigger an individual's movement behavior. The possibilities of actions triggered by (actual or even virtual) object features are called "affordances" (Gibson, 1979; Bub and Masson, 2010). In other words, affordances are *possibilities for action*. In the category of emotion and affection, vitality forms reflect the changing mood or affective state of the agent in an ongoing dyadic interaction, so it might inevitably be affected by the socialized contextualization of social affordances. One usage of the concept of social affordances literally means possibilities for intersubjective interactions. Another usage refers to affordances that are shaped by sociocultural context, as in Gibson's example of the postbox affordance of letter-mailing. It should be noted that one's perception of the postbox as affording letter-mailing is also determined by this person's previous sociocultural practice (imagine there is a country without a postal system) (cf. Gibson, 1979; de Carvalho, 2020; Borghi, 2021). In the first usage, social affordances can be seen as inherent properties of social surroundings that could constrain the possibilities of the actions executed by an agent. The second usage implies the malleable properties of social affordances.

For useful collections of existing resources, in the following passages of this subsection, actually we focus on the relevance of vitality forms and social affordances. In our view, both concepts in an essential way convey a sense of grounding relation between ground and its grounded things. The idea of social affordances implies a binary relation between affordances *per se* and their afforded (grounded) possible actions in a sociocultural context. In a similar way, vitality forms (as a ground) could also be regarded as *possibilities for action*, for the other two grounded higher dimensions of action, contents and intentions.[5]

Orban et al. (2021) advocate a promising neurophysiological framework of social cognition which can be firmly established on a rich source of social affordances of others.[6] Roughly, the importance of vitality forms is for communicating with others through performing the style of action. In this view, vitality forms are always expressed toward the other agent within a dyad. Regarding the influence of others and the role of vitality forms in social interaction, several studies demonstrate that gentle or rude actions performed by the agent, however, affect the motor response of the receiver (cf. Di Cesare et al., 2017a, 2021b; Lombardi et al., 2021). In general, both vitality forms and social affordances are related to the modality of action, so to speak, yet separately focusing on different aspects; the former is on the kinematic signatures of action, while the latter is on the possibility of action. Vitality forms highlight the kinematic nature of the action, i.e., the how-dimension of action; social affordances stress on the objective properties of surroundings within which the agent has to be constrained.

The presence of a conspecific requesting gesture will change the way an individual interacts with an object. It has been shown that social affordances are activated in interactions between conspecific individuals premised on the social intention of feeding, which changes the sequence kinematics of reach-grasp tasks and placement (Ferri et al., 2010). This is related to the same sequence of pointing to inanimate targets. Subsequent research has shown that the social request (i.e., the gesture of requesting to open the mouth) made by the recipient is a prerequisite for activating social affordances. Specifically, even if the sequence that points to the same individual does not finalize eating, the social request to be fed will activate social affordances. Moving around the space of the same species without making any social requests has little effect on the sequence. A conspecific gaze is a necessary condition for social requests to effectively activate social affordances. In general, the control of motor sequences can be altered by the interaction between the actor and the receiver; it is the characteristic of interaction that the actor activates social affordances based on the social requirements generated by the receiver. The gaze of the recipient is a prerequisite for the validity of the social request (Ferri et al., 2011). One of the most important human abilities is to understand the behavior of other conspecifics.

Another series of studies explore the effect of sudden demands on the kinematics of pre-planned actions (Sartori et al., 2009). In experiment 1, in 80% of the trials, participants are asked to grab an object and put it in a container (no interference test). In the other 20% of the trials, the assistant sitting next to the participant accidentally extends their arm and opens their hand, as if asking for the object (interference trial). In the remaining 3 experiments, (a) the assistant is replaced by a machine, (b) the gestures made by the assistant do not imply social demands, and (c) the assistant's gaze is unavailable. The results show that there is a kinematic change in the actions directed at the target only when the disturbance is a social request involving a human assistant. In contrast, there is no significant effect on kinematics when the interference is caused by a robot or by a human assistant performing a non-social gesture. These findings are discussed in the light of theories currently proposed to explain the influence of the social environment on action control. Another study aims to determine whether requested gestures and gaze direction are sufficient to infer communication intention in social contexts, by examining the influence of requested gestures and gaze direction on the kinematics of another individual's arm movements (Innocenti et al., 2012). Research shows that social requests activate social affordances, which interfere with the control of sequences, and that gaze from potential recipients who hold the cup in their hands modulates the effectiveness of gestures. This paradigm, when applied to individuals with autism, could provide new insights into the nature of their impairments in social interaction and communication.

Vitality forms as grounding social emotion

The category of emotion is different from that of cognition. In aiding the relevant discussions of vitality forms and their grounded social emotion, in this subsection, we turn to focus on a typical example of social emotion, that is, empathy. Empathy refers to the ability to put oneself in the situation of others in order to feel and understand both sides of feelings and cognitions. Vitality forms are not specific forms that could regulate the internal state of human movement behavior in a continuous manner. There might be a certain relationship between empathy and vitality forms. Twenty years ago, scholars propose the concept of the "mirror system," which is a significant

[5] We thank a referee for raising this issue.

[6] For discussions of orthodox Gibsonian social affordances (see Rietveld et al., 2019; Baggs, 2021); for applications of social affordances in the context of Bayesian brain modeling (see Ramstead et al., 2016; Veissière et al., 2020).

discovery (Rizzolatti et al., 1996; Gallese, 2001) in behavioral and social neuroscience. Its function is specifically manifested in that when the observer watches someone perform a goal-oriented action, the observer's mirror neurons would fire in the same pattern, just as if the observer themself is performing the action. It can also imply that the observer shares a virtual experience of that actor's experience. This has obvious meanings for understanding empathy and identification and response to artistic performance. How to capture the exact action characteristics of a specific individual in a certain way, and how to explain empathy and identification? Mirror neurons can suitably explain the "what" dimension of this behavior (goal-oriented). In addition, other mechanisms, such as intention detection centers, could explain the "why" dimension (intention and goal) (Ruby and Decety, 2001).

However, for recognition based on faithful imitation, people also need to understand the "how" dimension, the other person's "dynamic movement characteristics," i.e., their vitality forms (Hobson and Lee, 1999). Identification and internalization require greater complete immersion in the lively flowing experience of another person besides empathy. Without vitality forms, identification and internalization will act like rules of action rather than a state of perceived immersion. Heller and Haynal (1997) study therapeutic videotapes of high-risk patients who have repeated suicide attempts. The therapists could not predict who would attempt suicide again. Next, these two researchers and a panel of judges carefully examine the facial expressions of high-risk patients. Using Ekman and Friesen's Facial Action Coding System, they are unable to forecast who would attempt suicide again. However, when they examine the facial expressions of the therapists, they could make important predictions about which patients would attempt suicide. This shows from another perspective that people are unaware of the role of vitality forms in recognition and empathy, a process that involves empathy and counter-empathy.

Discussion

This article reviewed related currents on how to directly access other minds through vitality forms. Vitality forms are seen as a characterization of the action style based on the kinematics of actions, which might implicitly convey the internal state of the agent's feelings and intentions. Vitality forms play a highly crucial role in basic interpersonal interactions (especially the parental–infant dyad). The effective perception of vitality forms allows people to directly know the attitudes of others simply through their actions, and even obtain relevant information about non-existent objects to which the pretended actions are directed. Similarly, the proper expression of the vitality forms enables others to understand people's own internal state in the right way. The activation of the DCI when expressing and perceiving vitality forms strongly indicates that there is a neural mirroring mechanism of vitality forms (of actions) in the insula, which allows a person to express their own affective state and understand the affective state of others in the action.

However, current studies on vitality forms have focused on the identification of the characteristic dimension of mildness or rudeness of the action. It has shown that the idea of vitality forms involves the behavioral style of the action, so it does not only include mildness (or rudeness). In the single dimension of rudeness, actions can also be expressed as indecision or willingness, or cautiousness or recklessness. Future research perhaps could further expand and deepen this aspect. In addition, intention and behavioral style belong to two different dimensions of action. The brain mechanisms involved in perceiving action intention and vitality forms belong to two different areas. The former is the parietal frontal circuit, and the latter is the DCI. So, how can people perceive others' intentions at the same time when they perceive the vitality forms characteristics of other people's actions? What are the anatomical connections between the brain regions involved? Current studies are still forthcoming in this aspect. Finally, in the real-life social environment, actions are nested within the social context. In addition to arm movements, there are other important contextual cues, such as pre-existing information, the actor's facial expressions, or even the observer's interpersonal relationship with the actor. How do these contextual clues influence people to directly access other minds through the vitality forms of actions? Existing studies do not answer this question adequately, and this may be of great significance to the assessment of the perception of vitality forms of children with autism. After all, simple visual information may not be sufficient for children with autism to correctly encode vitality forms. Therefore, future studies might overcome this limitation by considering the use of different types of stimuli combinations.

Traditional achievements of cognitive neuroscience hold that action comprehension is closely related to the mirror neuron system (MNS). However, recent studies have shown that a greater number of brain regions are involved in action comprehension. As shown above, previous studies imply that there is a certain link between the neural bases of the vitality forms and the mirroring mechanism of neurons. The right DCI, the main brain region activated in perceiving and expressing vitality forms, may be endowed with a mirror mechanism that translates sensory information about other people's vitality forms into these forms of movement modules (Di Cesare et al., 2014, 2015). This may explain why children with ASD have difficulty in recognizing the vitality forms of other people's actions; i.e., they have difficulty relying on their own misidentification of the corresponding vitality forms when observing others' actions. It can also explain the difficulty of healthy children in recognizing the vitality forms of autistic children, that is, it is difficult for them to match the sensory modules of the motor kinematics observed in children with autism with their own brain processing and corresponding modules of vitality forms (see **Figure 1**). Di Cesare et al. (2021a) suggest that besides the DCI, the middle cingulate cortex

(MCC) is also strongly activated during action observation and execution. This new finding is involved with a new how-dimension of action, by using jerky movements as a control condition.

So, up to the present, we get two trustworthy neural bases of vitality forms: DCI and MCC.[7] In the end, in this article, for fulfilling the neural bases of vitality forms based on current neuroimaging studies by Rizzolatti, Di Cesare, and others, we

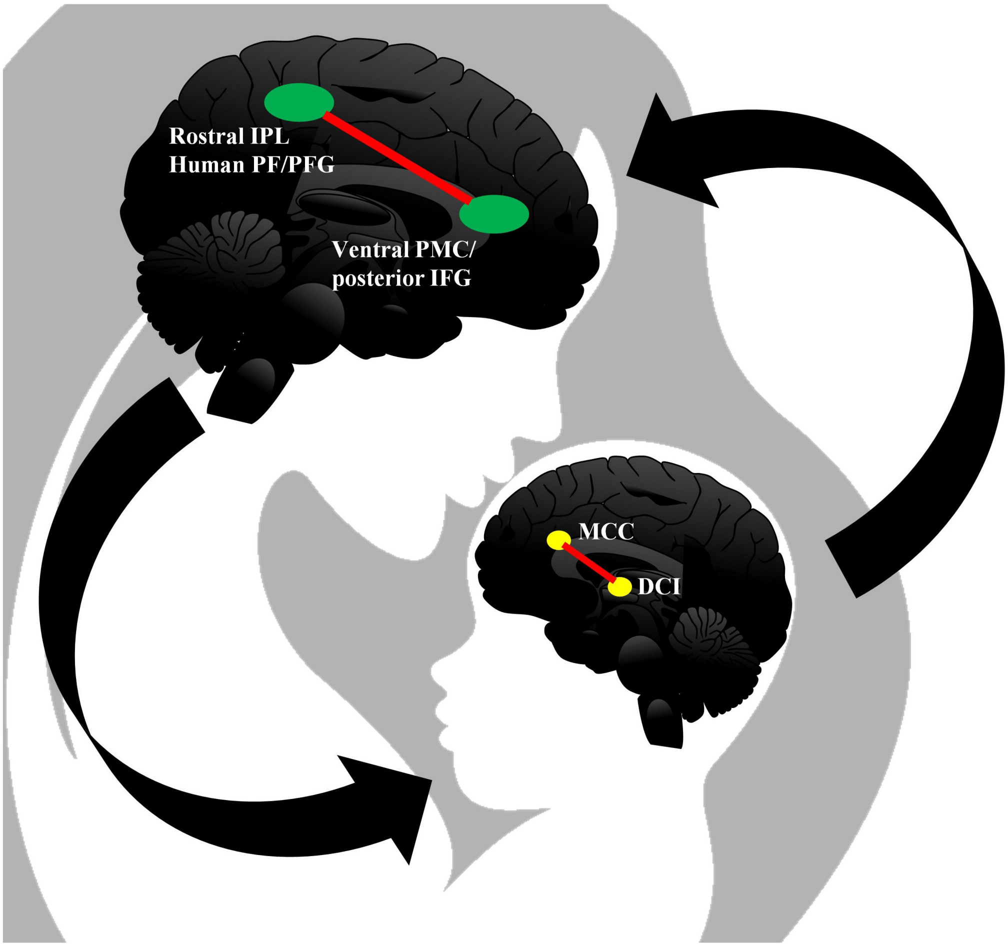


FIGURE 1
The neural bases of vitality forms (and human mirror neuron system, MNS) in a mother–infant communicative dyad. In the brain of the mother, the human MNS is composed of two parts: frontal MNS and parietal MNS. Generally, the anterior region with mirror neuron properties is located in the sub-frontal cortex, including the posterior frontal gyrus (IFG) and the adjacent ventral premotor cortex (PMC); the posterior region with mirror neuron properties is located in the rostral section of the inferior parietal lobe (IPL) and can be regarded as the human homolog of region PF/PFG in the macaque (cf. Pandya and Seltzer, 1982; Iacoboni and Mazziotta, 2007; Iacoboni, 2009; Bermúdez, 2020, p. 372). In the brain of the young baby, the picture presents a mirroring mechanism tuple (based on current trustworthy neuroimaging findings) of MCC and DCI, which can deal with observation and execution of vitality forms of action. Vitality forms occupy a new dimension apart from the what- and why-dimension of action, and ongoing actions *per se* convey enough in revealing the signatures of social perception. In a series of fMRI studies, researchers found selective activation of the DCI during both observation and execution of vitality forms (Di Cesare et al., 2014, 2017b). However, in one experiment, the MCC also showed activation (Di Cesare et al., 2020). In a subsequent fMRI study, the investigators employed the classical vitality forms paradigm, but eliminated the continuous style of movement by using jerky movements in a control condition to assess the role of the cingulate cortex in the processing of vitality forms. Participants performed two different tasks: observing a gentle or rude action and performing the same action. The results indicated that the MCC was strongly activated during action observation and execution, in addition to the insula (Di Cesare et al., 2021a). Besides, the neural bases of vitality forms should be incorporated into the larger MNS in the near future.

[7] Besides, another suggestion implies the potential role of the cerebellum (Ramsey et al., 2021). The role played by the cerebellum in social action comprehension is increasingly appreciated (Van Overwalle et al., 2019), and neurons with mirroring qualities are also present in the cerebellum (Molenberghs et al., 2012). Future research needs to elucidate the role of the cerebellum in social action comprehension. Individual action comprehension and learning may be closely related to mirror neurons, but the involvement of the cerebellum is essential when a series of actions are linked together in a temporal sequence. We imply that the relationship between the cerebellum and vitality forms is one of the future research directions.

highlight two areas, that is, DCI (actions executed in continuous condition) and MCC (actions executed in jerky condition), in Figure 1. Besides, the neural bases of vitality forms should be regarded as one part of the larger human MNS, since the former simply concerns the how-dimension of action rather than all dimensions involved in the social perception of dyadic interaction.

Author contributions

DD and WC: conceptualization and supervision. QL and JZ: investigation. QL, DD, and JZ: writing of the original draft. QL and DD: reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Social Science Fund of China (Grant Nos. 20CZX015 and 21BZX005).

Acknowledgments

We are thankful to the editor and three reviewers for their helpful comments on this manuscript.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th Edn. Washington, DC: American Psychiatric Association.

Ansuini, C., Cavallo, A., Campus, C., Quarona, D., Koul, A., and Becchio, C. (2016). Are we real when we fake? Attunement to object weight in natural and pantomimed grasping movements. *Front. Hum. Neurosci.* 10:471. doi: 10.3389/fnhum.2016.00471

Apperly, I. (2010). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Hove: Psychology Press.

Atkinson, A. P. (2009). Impaired recognition of emotions from body movements is associated with elevated motion coherence thresholds in autism spectrum disorders. *Neuropsychologia* 47, 3023–3029. doi: 10.1016/j.neuropsychologia.2009.05.019

Baggs, E. (2021). All affordances are social: Foundations of a Gibsonian social ontology. *Ecol. Psychol.* 33, 257–278. doi: 10.1080/10407413.2021.1965477

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Top. Cogn. Sci.* 2, 716–724. doi: 10.1111/j.1756-8765.2010.01115.x

Becchio, C., Koul, A., Ansuini, C., Bertone, C., and Cavallo, A. (2018). Seeing mental states: An experimental strategy for measuring the observability of other minds. *Phys. Life Rev.* 24, 67–80. doi: 10.1016/j.plrev.2017.10.002

Becchio, C., Sartori, L., Bulgheroni, M., and Castiello, U. (2008). The case of Dr. Jekyll and Mr. Hyde: A kinematic study on social intention. *Conscious. Cogn.* 17, 557–564. doi: 10.1016/j.concog.2007.03.003

Bermúdez, J. (2020). *Cognitive Science: An Introduction to the Science of the Mind*, 3rd Edn. New York, NY: Cambridge University Press.

Borghi, A. (2021). Affordances, context and sociality. *Synthese* 199, 12485–12515. doi: 10.1007/s11229-018-02044-1

Bub, D. N., and Masson, M. E. (2010). Grasping beer mugs: On the dynamics of alignment effects induced by handled objects. *J. Exp. Psychol. Hum. Percept. Perform.* 36, 341–358. doi: 10.1037/a0017606

Bystrom, K., Grahn, P., and Hagerhall, C. (2019). Vitality from experiences in nature and contact with animals—A way to develop joint attention and social engagement in children with autism? *Int. J. Environ. Res. Public Health* 16:4673. doi: 10.3390/ijerph16234673

Casartelli, L., Cesareo, A., Biffi, E., Campione, G., Villa, L., Molteni, M., et al. (2020a). Vitality form expression in autism. *Sci. Rep.* 10:17182. doi: 10.1038/s41598-020-73364-x

Casartelli, L., Federici, A., Fumagalli, L., Cesareo, A., Nicoli, M., Ronconi, L., et al. (2020b). Neurotypical individuals fail to understand action vitality form in children with autism spectrum disorder. *Proc. Natl. Acad. Sci. U.S.A.* 117, 27712–27718. doi: 10.1073/pnas.2011311117

Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., and Becchio, C. (2016). Decoding intentions from movement kinematics. *Sci. Rep.* 6:37036. doi: 10.1038/srep37036

Cavina-Pratesi, C., Kuhn, G., Ietswaart, M., and Milner, A. D. (2011). The magic grasp: Motor expertise in deception. *PLoS One.* 6:e16568. doi: 10.1371/journal.pone.0016568

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London: John Murray.

Davies, M., and Stone, T. (1995). *Folk Psychology: The Theory of Mind Debate*. New York, NY: John Wiley & Sons.

de Carvalho, E. M. (2020). "Social affordance," in *Encyclopedia of Animal Cognition and Behavior*, eds J. Vonk and T. Shackelford (Cham: Springer), 1–4.

Di Cesare, G., Cuccio, V., Marchi, M., Sciutti, A., and Rizzolatti, G. (2022). Communicative and affective components in processing auditory vitality forms: An fMRI study. *Cereb. Cortex* 32, 909–918. doi: 10.1093/cercor/bhab255

Di Cesare, G., De Stefani, E., Gentilucci, M., and De Marco, D. (2017a). Vitality forms expressed by others modulate our own motor response: A kinematic study. *Front. Hum. Neurosci.* 11:565. doi: 10.3389/fnhum.2017.00565

Di Cesare, G., Marchi, M., Errante, A., Fasano, F., and Rizzolatti, G. (2017b). Mirroring the social aspects of speech and actions: The role of the insula. *Cereb. Cortex* 28, 1348–1357. doi: 10.1093/cercor/bhx051

Di Cesare, G., Di Dio, C., Marchi, M., and Rizzolatti, G. (2015). Expressing our internal states and understanding those of others. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10331–10335. doi: 10.1073/pnas.1512133112

Di Cesare, G., Di Dio, C., Rochat, M. J., Sinigaglia, C., Bruschweiler-Stern, N., Stern, D. N., et al. (2014). The neural correlates of 'vitality form' recognition: An fMRI study: This work is dedicated to Daniel Stern, whose immeasurable contribution to science has inspired our research. *Soc. Cog. Affect. Neurosci.* 9, 951–960. doi: 10.1093/scan/nst068

Di Cesare, G., Fasano, F., Errante, A., Marchi, M., and Rizzolatti, G. (2016). Understanding the internal states of others by listening to action verbs. *Neuropsychologia* 89, 172–179. doi: 10.1016/j.neuropsychologia.2016.06.017

Di Cesare, G., Gerbella, M., and Rizzolatti, G. (2020). The neural bases of vitality forms. *Natl. Sci. Rev.* 7, 202–213. doi: 10.1093/nsr/nwz187

Di Cesare, G., Marchi, M., Lombardi, G., Gerbella, M., Sciutti, A., and Rizzolatti, G. (2021a). The middle cingulate cortex and dorso-central insula: A mirror circuit encoding observation and execution of vitality forms. *Proc. Natl. Acad. Sci. U.S.A.* 118:44. doi: 10.1073/pnas.2111358118

Di Cesare, G., Pelosi, A., Aresta, S., Lombardi, G., and Sciutti, A. (2021b). Affective contagion: How attitudes expressed by others influence our perception of actions. *Front. Hum. Neurosci.* 15:712550. doi: 10.3389/fnhum.2021.712550

Ferri, F., Campione, G. C., Dalla Volta, R., Gianelli, C., and Gentilucci, M. (2010). To me or to you? When the self is advantaged. *Exp. Brain Res.* 203, 637–646. doi: 10.1007/s00221-010-2271-x

Ferri, F., Campione, G. C., Dalla Volta, R., Gianelli, C., and Gentilucci, M. (2011). Social requests and social affordances: How they affect the kinematics of motor sequences during interactions between conspecifics. *PLoS One.* 6:e15855. doi: 10.1371/journal.pone.0015855

Gallagher, S. (2020). *Action and Interaction*. Oxford: Oxford University Press.

Gallagher, S., Sparaci, L., and Varga, S. (2022). Disruptions of the meshed architecture in autism spectrum disorder. *Psychoanal. Inq.* 42, 76–95. doi: 10.1080/07351690.2022.2007032

Gallese, V. (2001). The "shared manifold" hypothesis: From mirror neurons to empathy. *J. Conscious. Stud.* 8, 33–50. doi: 10.1159/000072786

Gallese, V., and Rochat, M. J. (2018). Forms of vitality: Their neural bases, their role in social cognition, and the case of autism spectrum disorder. *Psychoanal. Inq.* 38, 154–164. doi: 10.1080/07351690.2018.1405672

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York, NY: Oxford University Press.

Heller, M., and Haynal, V. (1997). "The doctor's face: Amirror of his patient's suicidal projects," in *The Body in Psychotherapy*, ed. J. Guimon (Basel, CH: Karger), 46–51.

Hobson, R. P., and Lee, A. (1999). Imitation and identification in autism. *J. Child Psychol. Psychiatry.* 40, 649–659. doi: 10.1111/1469-7610.00481

Hurley, S. (2008). Understanding simulation. *Philos. Phenomenol. Res.* 77, 755–774. doi: 10.1111/j.1933-1592.2008.00220.x

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annu. Rev. Psychol.* 60, 653–670. doi: 10.1146/annurev.psych.60.110707.163604

Iacoboni, M., and Mazziotta, J. C. (2007). Mirror neuron system: Basic findings and clinical applications. *Ann. Neurol.* 62, 213–218. doi: 10.1002/ana.21198

Innocenti, A., De Stefani, E., Bernardi, N. F., Campione, G. C., and Gentilucci, M. (2012). Gaze direction and request gesture in social interactions. *PLoS One.* 7:e36390. doi: 10.1371/journal.pone.0036390

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia* 33, 1419–1432. doi: 10.1016/0028-3932(95)00073-c

Kemmerer, D. (2021). What modulates the mirror neuron system during action observation?: Multiple factors involving the action, the actor, the observer, the relationship between actor and observer, and the context. *Prog. Neurobiol.* 205:102128. doi: 10.1016/j.pneurobio.2021.102128

Koul, A., Cavallo, A., Ansuini, C., and Becchio, C. (2016). Doing it your way: How individual movement styles affect action prediction. *PLoS One.* 11:e0165297. doi: 10.1371/journal.pone.0165297

Krueger, J. (2021). Enactivism, other minds, and mental disorders. *Synthese* 1, 365–389. doi: 10.1007/s11229-019-02133-9

Larsson, S. (2018). Grounding as a side-effect of grounding. *Top. Cogn. Sci.* 10, 389–408. doi: 10.1111/tops.12317

Lombardi, G., Zenzeri, J., Belgiovine, G., Vannucci, F., Rea, F., Sciutti, A., et al. (2021). The influence of vitality forms on action perception and motor response. *Sci. Rep.* 11:22576. doi: 10.1038/s41598-021-01924-w

Manera, V., Schouten, B., Becchio, C., Bara, B. G., and Verfaillie, K. (2010). Inferring intentions from biological motion: A stimulus set of point-light communicative interactions. *Behav. Res. Methods.* 42, 168–178. doi: 10.3758/BRM.42.1.168

Marraffa, M., and Meini, C. (2019). Forms of vitality revisited: The construction of an affective bodily self. *Theor. Psychol.* 29, 27–45. doi: 10.1177/0959354318822175

Molenberghs, P., Cunnington, R., and Mattingley, J. B. (2012). Brain regions with mirror properties: A meta-analysis of 125 human fMRI studies. *Neurosci. Biobehav. Rev.* 36, 341–349. doi: 10.1016/j.neubiorev.2011.07.004

Orban, G. A., Lanzilotto, M., and Bonini, L. (2021). From observed action identity to social affordances. *Trends Cogn. Sci* 25, 493–505. doi: 10.1016/j.tics.2021.02.012

Pandya, D. N., and Seltzer, B. (1982). Intrinsic connections and architectonics of posterior parietal cortex in the rhesus monkey. *J. Comp. Neurol.* 204, 196–210. doi: 10.1002/cne.902040208

Podda, J., Ansuini, C., Vastano, R., Cavallo, A., and Becchio, C. (2017). The heaviness of invisible objects: Predictive weight judgments from observed real and pantomimed grasps. *Cognition* 168, 140–145. doi: 10.1016/j.cognition.2017.06.023

Ramsey, R., Kaplan, D. M., and Cross, E. S. (2021). Watch and learn: The cognitive neuroscience of learning from others' actions. *Trends Neurosci.* 44, 478–491. doi: 10.1016/j.tins.2021.01.007

Ramstead, M. J. D., Veissière, S. P. L., and Kirmayer, L. J. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. *Front. Psychol.* 7:1090. doi: 10.3389/fpsyg.2016.01090

Rietveld, E., Rietveld, R., and Martens, J. (2019). Trusted strangers: Social affordances for social cohesion. *Phenomenol. Cogn. Sci.* 18, 299–316. doi: 10.1007/s11097-017-9554-7

Rizzolatti, G., D'Alessio, A., Marchi, M., and Di Cesare, G. (2021). The neural bases of tactile vitality forms and their modulation by social context. *Sci. Rep.* 11:9095. doi: 10.1038/s41598-021-87919-z

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cogn. Brain Res.* 3, 131–141. doi: 10.1016/0926-6410(95)00038-0

Rochat, M. J., and Gallese, V. (2022). The blurred vital contours of intersubjectivity in autism spectrum disorder: Early signs and neurophysiological hypotheses. *Psychoanal. Inq.* 42, 30–52. doi: 10.1080/07351690.2022.2007022

Rochat, M. J., Veroni, V., Bruschweiler-Stern, N., Pieraccini, C., Bonnet-Brilhault, F., Barthélémy, C., et al. (2013). Impaired vitality form recognition in autism. *Neuropsychologia* 51, 1918–1924. doi: 10.1016/j.neuropsychologia.2013.06.002

Ruby, P., and Decety, J. (2001). Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Natl. Neurosci.* 4, 546–550. doi: 10.1038/87510

Sartori, L., Becchio, C., Bulgheroni, M., and Castiello, U. (2009). Modulation of the action control system by social intention: Unexpected social requests override preplanned action. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1490. doi: 10.1037/a0015777

Sartori, L., Becchio, C., and Castiello, U. (2011). Cues to intention: The role of movement information. *Cognition.* 119, 242–252. doi: 10.1016/j.cognition.2011.01.014

Stern, D. N. (1977). *The First Relationship: Infant and Mother*. Cambridge, NY: Harvard University Press.

Stern, D. N. (1998). *The Interpersonal World of the Infant: A View from Psychoanalysis and Developmental Psychology*. London: Karnac Books.

Stern, D. N. (2010). *Forms of Vitality: Exploring Dynamic Experience in Psychology, the Arts, Psychotherapy, and Development*. New York, NY: Oxford University Press.

Trevarthen, C. (2013). Dan Stern's voyage of discovery in the interpersonal world of human movement, and the gifts he brought to us. *Self Soc.* 40, 42–45. doi: 10.1080/03060497.2013.11084283

Trevarthen, C. (2019). Sander's life work, on mother-infant vitality and the emerging person. *Psychoanal. Inq.* 39, 22–35. doi: 10.1080/07351690.2019.1549909

Van Overwalle, F., De Coninck, S., Heleven, E., Perrotta, G., Taib, N. O. B., Manto, M., et al. (2019). The role of the cerebellum in reconstructing social action sequences: A pilot study. *Soc. Cogn. Affect. Neurosci.* 14, 549–558. doi: 10.1093/scan/nsz032

Veissière, S., Constant, A., Ramstead, M., Friston, K., and Kirmayer, L. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behav. Brain Sci.* 43:E90. doi: 10.1017/S0140525X19001213

frontiers | Frontiers in Psychology

TYPE Mini Review
PUBLISHED 22 September 2022
DOI 10.3389/fpsyg.2022.968382

Check for updates

OPEN ACCESS

EDITED BY
Luca Tummolini,
National Research Council (CNR), Italy

REVIEWED BY
Michał Klincewicz,
Tilburg University,
Netherlands
Rudolf M. Oosthuizen,
University of South Africa, South Africa

*CORRESPONDENCE
Christian U. Krägeloh
chris.krageloh@aut.ac.nz

SPECIALTY SECTION
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

RECEIVED 13 June 2022
ACCEPTED 29 August 2022
PUBLISHED 22 September 2022

CITATION
Krägeloh CU, Bharatharaj J, Albo-Canals J, Hannon D and Heerink M (2022) The time is ripe for robopsychology. *Front. Psychol.* 13:968382. doi: 10.3389/fpsyg.2022.968382

COPYRIGHT
© 2022 Krägeloh, Bharatharaj, Albo-Canals, Hannon and Heerink. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The time is ripe for robopsychology

Christian U. Krägeloh[1]*, Jaishankar Bharatharaj[2], Jordi Albo-Canals[3], Daniel Hannon[4] and Marcel Heerink[5]

[1]PAIR Lab New Zealand, Auckland University of Technology, Auckland, New Zealand, [2]PAIR Lab India, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India, [3]Lighthouse Disruptive Innovation Group, LLC., Cambridge, MA, United States, [4]Department of Mechanical Engineering, Tufts University, Medford, MA, United States, [5]Saxion University of Applied Sciences, Enschede, Netherlands

As robotic applications become increasingly diverse, more domains of human lives are being involved, now also extending to educational, therapeutic, and social situations, with a trend to even more complex interactions. This diversity generates new research questions that need to be met with an adequate infrastructure of psychological methods and theory. In this review, we illustrate the current lack of a sub-discipline in psychology to systematically study the psychological corollaries of living in societies where the application of robotic and artificial intelligence (AI) technologies is becoming increasingly common. We thus propose that organized efforts be made toward recognition of robopsychology as a sub-discipline so that the field of psychology moves away from isolated publications of robot- and AI-related topics to a body of knowledge that is able to meet the demands for change, as the world is preparing for the Fourth Industrial Revolution. We propose a definition of robopsychology that not only covers the study of the effects of robots on human behavior, but also of robots and AI themselves, as well as acknowledging how this sub-discipline may eventually be fundamentally changed through robots and AI. In this sense, our definition mirrors an already existing definition of the field of robophilosophy.

KEYWORDS

psychology, sub-discipline, special interest group, robot psychology, robotic psychology, robopsychology, robots, artificial intelligence

Introduction

While the word *robot* has only first appeared in the early 1920s through Karel Čapek's science fiction play *R.U.R.* (Rossum's Universal Robots; Čapek, 2004), the idea of self-moving machines, or automata, has featured in myths and stories that go back three millennia and are found in many parts of the world (Mayor, 2018). Throughout history, there have also been many independent attempts to create actual automata, such as water-powered organs or mechanized beasts and androids (Cave and Dihal, 2018). Of course, it was not until the rapid technological progress of the 20th century that robots became more wide spread. For example, the adoption of robotic technology in the automotive manufacturing industry resulted in dramatic increases in cost-efficiency and production quality (Karabegović, 2016). Due to the precision they provide, robots have also become commonplace in medical contexts such as in surgery (Lane, 2018).

Most individuals would rarely encounter industrial and surgical robots, and if so, only witness the very specific functions that these robots provide. This is in contrast with the notion of robots as embodied intelligent and autonomous agents and particularly with portrayal in media and film, where robots often appear as highly sophisticated and with the potential to lead to utopian or dystopian scenarios (Mubin et al., 2019). Fact-based media reporting has been shown to increase positive attitudes and trust in robots (Savela et al., 2021), and actual encounters with robots also have the potential to alleviate much of the anxiety and wariness that people may have. While instances of robotic hotel check-in and room service (Fuentes-Moraleda et al., 2020) or robotic chefs in restaurants (Fusté-Forné, 2021) may still be viewed as having primarily entertainment value, systematic attempts have increasingly been made to apply robots to provide psychosocial or educational benefits for humans. Robots have thus been used to provide companionship for older people (Gasteiger et al., 2021), robot-enhanced psychotherapy (Costescu et al., 2014), or to assist in learning and teaching (Belpaeme et al., 2018). As human-robot interactions are appearing to become more lively and reciprocal, more effort is directed at studying the psychological reactions of human users in order to optimize this experience. Research has thus explored the effects of a range of variables such as robot morphology (Mara et al., 2022), voice (Dong et al., 2021), or nonverbal behavior (Zinina et al., 2020). Recent research has even explored the extent to which what the appearance of robots may be racialized with the potential to perpetuate racial stereotypes (Bartneck et al., 2018).

The trend towards increased relevance of robots in people's lives accelerates the need to understand the variables that influence the quality of human-robot interactions as well as their psychological corollaries. While a solid body of research has already emerged (Siciliano and Khatib, 2016), new research questions continue to be posed, particularly the extent to which such applications are motivated by or fulfil humans' psychological needs. As shown by robotic pets (Melson et al., 2009), robotic romance (Viik, 2020), sex robots (Döring et al., 2020), or robots to provide spiritual and religious support (Trovato et al., 2021), human-robot interactions are increasing in complexity, thus connecting robot research with the same psychological models and theories that are used to explain social behaviors among humans, such as attachment theory (Pozharliev et al., 2021) or social identity theory (Edwards et al., 2019). The purpose of the present review was to explore the extent to which there are any existing sub-disciplines in psychology devoted to the study of topics involving robots. Using a state-of-the-art review approach (Grant and Booth, 2009) with a systematic search strategy, we provided an outline of the landscape of psychological sub-disciplines. Not being based on any previous theories or hypotheses, this review followed an inductive approach (Watson et al., 2018).

A review of the representation of the study of robots in existing sub-disciplines of psychology

At the time of writing this review (May 2022), the journal *Frontiers in Psychology* listed 32 sub-disciplines or sub-fields of psychology to structure the content of its articles (Frontiers in Psychology, 2022). We present these in Table 1, together with potential sub-discipline names expressed through the 54 divisions recognized by the American Psychological Association (APA) at the time of writing this review (APA, 2022a). APA notes that some of the divisions represent special interest groups rather than sub-disciplines. However, for the purposes of identifying representation of robotics-related research in psychology, including special interest groups in addition to sub-disciplines provides a more comprehensive analysis. Additionally, we searched through the APA literature database PsycInfo for journal titles that could indicate a sub-discipline that may have recently emerged or is too small to have been recognized yet as a sub-field in psychology. We searched this psychology database containing nearly 2,300 journals for the word stem "psycho" to identify potential sub-discipline names that are expressed either by a preceding adjective other than a geographical location (e.g., applied psychology), a preceding noun (e.g., community psychology), or by a prefix (e.g., ecopsychology). The presence of two adjectives was considered to be too specific and indicative of a further sub-categorization within a sub-discipline. For example, *applied social psychology* was not included as it was treated as a further division of *social psychology*. If a name contained two adjectives (e.g., *reproductive and infant psychology*), the entry was presented like that, unless both adjectives had already resulted in a separate entry. Synonyms or very similar terms were still retained as separate entries, such that both *child psychology* and *pediatric psychology* were included. The search was conducted by the first author using coding for relevance, which was verified independently by the second author. Any uncertainty was resolved by discussion. In total, we list 127 entries in Table 1, with information on where they were sourced from.

None of the 127 entries in Table 1 make any reference to robots. APA Division 21 (*Applied Experimental and Engineering Psychology*) might initially appear to have some relevance to robotics but is very broadly worded as promoting "the development and application of psychological principles, knowledge, and research to improve technology, consumer products, energy systems, communication and information, transportation, decision making, work settings and living environments" (APA, 2022b). While three journal titles in PsycInfo contained the word "robot," none of these are representative of what may be considered a relevant sub-discipline of psychology. *ACM Transaction of Human-Robot Interaction* is described on its homepage (Association for Computing Machinery, 2022) to be an interdisciplinary journal that also welcomes submissions from behavioral and social sciences. *Intelligent Service Robotics* (Springer, 2022a) is focused on assistive

TABLE 1 Sub-disciplines and special interest groups of psychology as presented by *Frontiers in Psychology*, *APA*, and in academic journal titles.

Sub-discipline	Source
Addiction Psychology	APA Division 50
Advancement of Psychotherapy	APA Division 29
Aerospace Psychology	Journal name "The International Journal of Aerospace Psychology"
Aging Psychology	Journal name "Aging Psychology"
American Psychology-Law Society	APA Division 41
Analytical Psychology	Journal name "The Journal of Analytical Psychology"
Animal Psychology	Journal name "Japanese Journal of Animal Psychology"
Auditory Cognitive Neuroscience	Frontiers in Psychology section
Adult Development and Aging	APA Division 20
Applied Experimental and Engineering Psychology	APA Division 21
Applied Psychology	Journal name "Journal of Applied Psychology"
Aviation Psychology	Journal name "Aviation Psychology and Applied Human Factors"
Behavioral Neuroscience and Comparative Psychology	APA Division 6
Behavioral Psychology	Journal name "Behavioral Psychology"
Behavior Analysis	APA Division 25
Biological Psychology	Journal name "Biological Psychology"
Black Psychology	Journal name "Journal of Black Psychology"
Child and Family Policy and Practice	APA Division 37
Child and Adolescent Psychology	Journal name "Journal of Clinical Child and Adolescent Psychology"
Child Psychology	Journal name "Educational and Child Psychology"
Clinical Child and Adolescent Psychology	APA Division 53
Clinical Neuropsychology	APA Division 40
Clinical Psychology	APA Division 12
Coaching Psychology	Journal name "International Coaching Psychology Review"
Cognition	Frontiers in Psychology section
Cognitive Psychology	Journal name "Cognitive Psychology"
Cognitive Science	Frontiers in Psychology section
Community Psychology	APA Division 27
Comparative Psychology	Frontiers in Psychology section
Consciousness Research	Frontiers in Psychology section
Constructivist Psychology	Journal name "Journal of Constructivist Psychology"
Counseling Psychology	APA Division 17
Consulting Psychology	APA Division 13
Consumer Psychology	APA Division 23
Couple and Family Psychology	APA Division 43
Cross-Cultural Psychology	Journal name "Journal of Cross-Cultural Psychology"
Cultural Psychology	Frontiers in Psychology section
Cyberpsychology	Journal name "Cyberpsychology, Behavior, and Social Networking"
Decision Neuroscience	Frontiers in Psychology section
Developmental Psychology	APA Division 7; Frontiers in Psychology section
Eating Behavior	Frontiers in Psychology section
Ecological Psychology	Journal name "Ecological Psychology"
Economic Psychology	Journal name "Journal of Economic Psychology"
Ecopsychology	Journal name "Ecopsychology"
Educational Psychology	APA Division 15; Frontiers in Psychology section
Emotion Science	Frontiers in Psychology section
Environmental, Population and Conservation Psychology	APA Division 34
Environmental Psychology	Frontiers in Psychology section
Ethnic Minority Psychology	Journal name "Cultural Diversity and Ethnic Minority Psychology"
Evolutionary Psychology	Frontiers in Psychology section

(Continued)

TABLE 1 Continued

Sub-discipline	Source
Experimental Psychology and Cognitive Science	APA Division 3
Family Psychology	Journal name “Journal of Family Psychology”
Forensic and Legal Psychology	Frontiers in Psychology section
Forensic Psychology	Journal name “American Journal of Forensic Psychology”
Gender, Sex and Sexualities	Frontiers in Psychology section
General Psychology	APA Division 1
Genetic Psychology	Journal name “The Journal of Genetic Psychology: Research and Theory on Human Development”
Gerontopsychology	Journal name “GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry”
Group Psychology and Group Psychotherapy	APA Division 49
Health Psychology	APA Division 38; Frontiers in Psychology section
Health Service Psychology	Journal name “Journal of Health Service Psychology: An Official Journal of the National Register of Health Service Psychologists”
History of Psychology	APA Division 26
Humanistic Psychology	APA Division 32
Human-Media Interaction	Frontiers in Psychology section
Individual Psychology	Journal name “The Journal of Individual Psychology”
Industrial and Organizational Psychology	APA Division 14
Intellectual and Developmental Disabilities/Autism Spectrum Disorder	APA Division 33
International Psychology	APA Division 52
Investigative Psychology	Journal name “Journal of Investigative Psychology and Offender Profiling”
Language Sciences	Frontiers in Psychology section
Latinx Psychology	Journal name “Journal of Latinx Psychology”
Legal and Criminological Psychology	Journal name “Legal and Criminological Psychology”
Managerial Psychology	Journal name “Journal of Managerial Psychology”
Mathematical Psychology	Journal name “Journal of Mathematical Psychology”
Mathematical and Statistical Psychology	Journal name “British Journal of Mathematical and Statistical Psychology”
Media Psychology and Technology	APA Division 46
Medical Psychology	Journal name “Medizinische Psychologie” [German]
Military Psychology	APA Division 19
Movement Science and Sport Psychology	Frontiers in Psychology section
Neuropsychology	Frontiers in Psychology section
Occupational and Organizational Psychology	Journal name “Journal of Occupational and Organizational Psychology”
Organizational Psychology	Frontiers in Psychology section
Pastoral Psychology	Journal name “Pastoral Psychology”
Peace Psychology	APA Division 48
Pediatric Psychology	APA Division 54; Frontiers in Psychology section
Perception Science	Frontiers in Psychology section
Performance Science	Frontiers in Psychology section
Personality and Social Psychology	APA Division 8; Frontiers in Psychology section
Personnel Psychology	Journal name “Personnel Psychology”
Phenomenological Psychology	Journal name “Journal of Phenomenological Psychology”
Philosophical Psychology	Journal name “Philosophical Psychology”
Police and Criminal Psychology	Journal name “Journal of Police and Criminal Psychology”
Political Psychology	Journal name “Political Psychology”
Positive Psychology	Frontiers in Psychology section
Prescribing Psychology	APA Division 55
Professional Psychology	Journal name “Professional Psychology: Research and Practice”
Projective Psychology	Journal name “Journal of Projective Psychology & Mental Health”

(Continued)

TABLE 1 Continued

Sub-discipline	Source
Psychoanalysis and Psychoanalytic Psychology	APA Division 39
Psychological Hypnosis	APA Division 30
Psychological Study of Culture, Ethnicity and Race	APA Division 45
Psychological Study of Men and Masculinities	APA Division 51
Psychological Study of Social Issues	APA Division 9
Psychologists in Independent Practice	APA Division 42
Psychologists in Public Service	APA Division 18
Psychology for Clinical Settings	Frontiers in Psychology section
Psychology of Aging	Frontiers in Psychology section
Psychology of Aesthetics, Creativity and the Arts	APA Division 10
Psychology of Religion and Spirituality	APA Division 36
Psychology of Sexual Orientation and Gender Diversity	APA Division 44
Psychology of Women	APA Division 35
Psycho-Oncology	Frontiers in Psychology section
Psychopharmacology and Substance Abuse	APA Division 28
Psychopathology	Frontiers in Psychology section
Qualitative Psychology	Journal name "Qualitative Psychology"
Quantitative and Qualitative Methods	APA Division 5
Quantitative Psychology and Measurement	Frontiers in Psychology section
Reading Psychology	Journal name "Reading Psychology"
Rehabilitation Psychology	APA Division 22
Reproductive and Infant Psychology	Journal name "Journal of Reproductive and Infant Psychology"
School Psychology	APA Division 16
Social Psychology	Journal name "Social Psychology"
Sport, Exercise and Performance Psychology	APA Division 47
State, Provincial and Territorial Psychological Association Affairs	APA Division 31
Teaching of Psychology	APA Division 2
Theoretical and Philosophical Psychology	APA Division 24; Frontiers in Psychology section
Transpersonal Psychology	Journal name "Journal of Transpersonal Psychology"
Trauma Psychology	APA Division 56

The entries are listed in alphabetical order. For journal titles, representative examples are shown.

functions of robots, making some mention of the relevance of cognitive science, and *International Journal of Social Robotics* (Springer, 2022b) is presented as an interdisciplinary journal that does not mention psychology specifically.

Discussion: Robot psychology, robotic psychology, or robopsychology?

The list in Table 1 indicates that there is currently no sub-discipline in psychology that can be considered to be giving robots special attention, either as experimental subjects or by studying their effects on human behavior. Of course, this does not mean that a potential psychological sub-discipline may not already have some sort of presence in the literature through individual publications. What are some potential sub-discipline names mentioned in this work and what do these names suggest about the way in which robots are studied? When searching the academic literature (using GoogleScholar) for "robot psychology," a small number of articles can be found. This includes a technical note by Konolige (1985) where *experimental robot psychology* is purported to be about "analyzing the design of a robot agent's cognitive processes" (p. 2). Gallagher (2013) referred to robot psychology when describing a robot's understanding of its own propositional attitudes (as equivalent to folk psychology for humans), and Nitsch and Popp (2014) used the term in the context of describing how robots as social agents need to be able to "predict human intentions and actions and display behavior that is appropriate to that context" (p. 622). Therefore, just like animal psychology is about understanding the behavior of animals, robot psychology is focused on robots only and thus not aspects related to the human perspective when interacting with robots.

A suitable alternative to *robot psychology* is *robotic psychology*. While this phrase has also been mentioned only very little in the literature, it has been clearly defined as the study of "individual differences in people's interactions with various robots, as well as the diversity of the robots themselves, applying principles of differential psychology to the traditional fields of human factors and human–computer interactions" (Libin and Libin, 2004, p. 1792). The authors contrasted robotic psychology with *robopsychology*, which they defined as "a systematic study of compatibility between people and artificial creatures" as well as the study of "psychological mechanisms of the animation of the technological entity which result in a unique phenomenon defined as a robot's 'personality'" (p. 1792). Unlike robotic psychology, which "focuses on the psychological significance of person-robotic creature communication" (Libin and Libin, 2004, p. 1792), the focus of robopsychology is thus on the understanding of robot behavior. This usage of the term is consistent with how it was first used when introduced as the name of a fictional science in short stories by Isaac Asimov in 1950 (Bátfai, 2020).

While some studies (Servick, 2019) have interpreted the term robopsychology in a way consistent with the definition above, other researchers have used the term interchangeably with robotic psychology (Duradoni et al., 2021; Linz Institute of Technology, 2022). In the absence of any well-established or consistent use of any of these terms, a future sub-discipline in psychology related to robots may still decide on a suitable name. In our view, the term *robopsychology* is preferable as it can be easily identified alongside the already established field of *robophilosophy* (Tzafestas, 2016) – the "philosophy *of*, *for*, and *by* social robotics" (Seibt, 2018, p.390). Philosophy *of* social robotics is seen as the reflective activities about conceptual implications of investigating human-robot interactions, while philosophy *for* reflects on conceptual norms, sociality, human capacities, social roles as well as legal and ethical responsibilities, and philosophy *by* expresses any fundamental re-orientation of philosophical research that might occur due to its activities (Seibt, 2018).

The tentative definition of robopsychology that we would like to offer is similar: the psychology *of*, *for*, and *by* robots, robotics, and artificial intelligence (AI). This wording contains a broader scope than social robots only. Additionally, *robots and robotics* expresses the fact that both the actual products as well as the ongoing process of designing and building robots are worthy topics for psychological research. We also propose to add AI so that the sub-discipline is not only limited to physical manifestations but also considers latent processes related to this technology. In this definition, the psychology *of* robots, robotics, and AI addresses psychological implications of encountering robots and AI as well as people's views regarding this technology. Psychology *for* concerns areas that are relevant in the design of robots and AI and the facilitation of the robotic applications in society. Lastly, psychology *by* acknowledges any fundamental changes in the way in which psychological topics in the study of robots and AI may be approached in the future. The latter can include issues such as transhumanism (DeFalco, 2020) and expresses the potential for AI to eventually even participate in the discipline of psychology.

Conclusion: The need for a science of robopsychology

With the predicted arrival of the so-called Fourth Industrial Revolution characterized by transformation through robotics and automation (Karabegović et al., 2020), psychological research can be expected to experience transformational changes. A rapidly expanding scope of application of robotic technology is already noticeable as robotics has moved from primarily industrial uses to areas involving direct contact with people, such as robots in the service industry, in educational settings, and as social agents. As our review illustrated, there is currently no psychological sub-discipline dedicated to the study of the effects that robots have on people's lives, which is currently only addressed through interdisciplinary fields such as human-robot interaction or social robotics. The advantages of organizing psychological research through the formation of special interest groups and sub-disciplines is undoubtedly the driver of the richness and diversity demonstrated in Table 1 of our review. With this review, we encourage activities toward the recognition of robopsychology as the sub-discipline that enables the necessary academic and theoretical infrastructure to facilitate psychological investigations in this changing world. Such work requires specific psychological theories and models to describe the increasing complexities of human interactions with robots, such as intimacy and spirituality, as well as suitable research methods and measurement of psychological constructs that meet quality standards for psychological research (Krägeloh et al., 2019). Our proposed definition of robopsychology is intentionally broad to permit a range of future applications and may be considered parallel to the already existing sub-discipline of robophilosophy. To what extent there is eventual demand for the sub-discipline of robopsychology is up for the future to decide. With this article, we hope to instigate the necessary debates.

Author contributions

JB conceived of the idea of proposing the field of robopsychology, which was subsequently discussed by all authors. CK created the proposed definition for the field of robopsychology, conducted the review, and provided the first draft. All authors contributed to the article and approved the submitted version.

Conflict of interest

JA-C was employed by the company Lighthouse Disruptive Innovation Group, LLC.

The remaining 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

American Psychological Association (APA) (2022a). APA Division. Available at: https://www.apa.org/about/division (Accessed June 6, 2022).

American Psychological Association (APA) (2022b). Applied Experimental and Engineering Psychology. Available at: https://www.apa.org/about/division/div21 (Accessed June 6, 2022).

Association for Computing Machinery (2022). ACM Transactions on Human-Robot Interaction. Available at: https://dl.acm.org/journal/thri (Accessed June 6, 2022).

Bartneck, C., Yogeeswaran, K., Ser, Q. M., Woodward, G., Sparrow, R., Wang, S., et al. (2018). "Robots and racism," in *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction,* 196–204.

Bátfai, N. (2020). Hacking with God: a common programming language of robopsychology and robophilosophy. arXiv preprint arXiv:2009.09068. doi: 10.48550/arXiv.2009.09068

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: a review. *Sci. Robot.* 3:aat5954. doi: 10.1126/scirobotics.aat5954

Čapek, K. (2004). *R.U.R (Rossum's Universal Robots)*. London: Penguin.

Cave, S., and Dihal, K. (2018). Ancient dreams of intelligent machines: 3,000 years of robots. *Nature* 559, 473–475. doi: 10.1038/d41586-018-05773-y

Costescu, C. A., Vanderborght, B., and David, D. O. (2014). The effects of robot-enhanced psychotherapy: a meta-analysis. *Rev. Gen. Psychol.* 18, 127–136. doi: 10.1037/gpr0000007

DeFalco, A. (2020). Towards a theory of posthuman care: real humans and caring robots. *Body Soc.* 26, 31–60. doi: 10.1177/1357034X20917450

Dong, J., Lawson, E., Olsen, J., and Jeon, M. (2021). Female voice agents in fully autonomous vehicles are not only more likeable and comfortable, but also more competent. *Proc. Hum. Factors Ergon. Soc.* 64, 1033–1037. doi: 10.1177/1071181320641248

Döring, N., Mohseni, M. R., and Walter, R. (2020). Design, use, and effects of sex dolls and sex robots: scoping review. *J. Med. Internet Res.* 22:e18551. doi: 10.2196/18551

Duradoni, M., Colombini, G., Russo, P. A., and Guazzini, A. (2021). Robotic psychology: a PRISMA systematic review on social-robot-based interventions in psychological domains. *Journals* 4, 664–709. doi: 10.3390/j4040048

Edwards, C., Edwards, A., Stoll, B., Lin, X., and Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: social identity theory in human-robot interactions. *Comput. Hum. Behav.* 90, 357–362. doi: 10.1016/j.chb.2018.08.027

Frontiers in Psychology (2022). Frontiers in Psychology - Sections. Available at: https://www.frontiersin.org/journals/psychology (Accessed June 6, 2022).

Fuentes-Moraleda, L., Díaz-Pérez, P., Orea-Giner, A., Muñoz-Mazónc, A., and Villacé-Molinero, T. (2020). Interaction between hotel service robots and humans: a hotel-specific service robot acceptance model (sRAM). *Tour. Manag. Perspect.* 36:100751. doi: 10.1016/j.tmp.2020.100751

Fusté-Forné, F. (2021). Robot chefs in gastronomy tourism: what's on the menu? *Tour. Manag. Perspect.* 37:100774. doi: 10.1016/j.tmp.2020.100774

Gallagher, S. (2013). You and I, robot. *AI & Soc.* 28, 455–460. doi: 10.1007/s00146-012-0420-4

Gasteiger, N., Loveys, K., Law, M., and Broadbent, E. (2021). Friends from the future: a scoping review of research into robots and computer agents to combat loneliness in older people. *Clin. Interv. Aging* 16, 941–971. doi: 10.2147/CIA.S282709

Grant, M. J., and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info. Libr. J.* 26, 91–108. doi: 10.1111/j.1471-1842.2009.00848.x

Karabegović, I. (2016). The role of industrial robots in the development of automotive industry in China. *Int. J. Eng. Works* 3, 92–97.

Karabegović, I., Turmanidze, R., and Dašić, P. (2020). "Robotics and automation as a foundation of the fourth industrial revolution-industry 4.0," in *Advanced Manufacturing Processes. Inter Partner 2019. Lecture Notes in Mechanical Engineering*. eds. V. Tonkonogyi, V. Ivanov, J. Trojanowska, G. Oborskyi, M. Edl, I. Kuric, et al. (Cham: Springer).

Konolige, K. G. (1985). *Experimental Robot Psychology*. Washington, DC: SRI International.

Krägeloh, C. U., Bharatharaj, J., Kutty, S. K. S., Nirmala, P. R., and Huang, L. (2019). Questionnaires to measure acceptability of social robots: a critical review. *Robotics* 8:88. doi: 10.3390/robotics8040088

Lane, T. (2018). A short history of robotic surgery. *Ann. R. Coll. Surg. Engl.* 100, 5–7. doi: 10.1308/rcsann.supp1.5

Libin, A. V., and Libin, E. V. (2004). Person-robot interactions from the robopsychologists' point of view: the robotic psychology and robotherapy approach. *Proc. IEEE* 92, 1789–1803. doi: 10.1109/JPROC.2004.835366

Linz Institute of Technology (2022). Robosychology Lab am Linz Institute of Technology. Available at: https://www.jku.at/lit-robopsychology-lab (Accessed June 6, 2022).

Mara, M., Appel, M., and Gnambs, T. (2022). Human-like robots and the uncanny valley: a meta-analysis of user responses based on the Godspeed scales. *Z. Psychol.* 230, 33–46. doi: 10.1027/2151-2604/a000486

Mayor, A. (2018). *Gods and robots – myths, machines, and ancient dreams of technology*. Princeton, NJ: Princeton University Press.

Melson, G. F., Kahn, P. H., Beck, A., and Friedman, B. (2009). Robotic pets in human lives: implications for the human–animal bond and for human relationships with personified technologies. *J. Soc. Issues* 65, 545–567. doi: 10.1111/j.1540-4560.2009.01613.x

Mubin, O., Wadibhasme, K., Jordan, P., and Obaid, M. (2019). Reflecting on the presence of science fiction robots in computing literature. *ACM Trans. Hum.-Robot Interact.* 8, 1–25. doi: 10.1145/3303706

Nitsch, V., and Popp, M. (2014). Emotions in robot psychology. *Biol. Cybern.* 108, 621–629. doi: 10.1007/s00422-014-0594-6

Pozharliev, R., De Angelis, M., Rossi, D., Romani, S., Verbeke, W., and Cherubino, P. (2021). Attachment styles moderate customer responses to frontline service robots: evidence from affective, attitudinal, and behavioral measures. *Psychol. Mark.* 38, 881–895. doi: 10.1002/mar.21475

Savela, N., Turja, T., Latikka, R., and Oksanen, A. (2021). Media effects on the perceptions of robots. *Hum. Behav. Emerg. Technol.* 3, 989–1003. doi: 10.1002/hbe2.296

Seibt, J. (2018). "Robophilosophy," in *Posthuman Glossary*. eds. R. Braidotti and M. Hlavajova (London: Bloomsbury), 390–393.

Servick, K. (2019). Could robots be psychology's new lab rats? *Science*. doi: 10.1126/science.aaz7641

Siciliano, B., and Khatib, O. (2016). *Springer Handbook of Robotics*. New York: Springer.

Springer (2022a). Intelligent Service Robotics. Available at: https://springer.com/journal/11370 (Accessed June 6, 2022).

Springer (2022b). International Journal of Social Robotics. Available at: https://springer.com/journal/12369 (Accessed June 6, 2022).

Trovato, G., De Saint Chamas, L., Nishimura, M., Paredes, R., Lucho, C., Huerta-Mercado, A., et al. (2021). Religion and robots: towards the synthesis of two extremes. *Int. J. Soc. Robot.* 13, 539–556. doi: 10.1007/s12369-019-00553-8

Tzafestas, S. G. (2016). *An Introduction to Robophilosophy: Cognition, Intelligence, Autonomy, Consciousness, Conscience and Ethics*. Sterling, VA: Stylus Publishing.

Viik, T. (2020). Falling in love with robots: a phenomenological study of experiencing technological alterities. *Paladyn, J. Behav. Robot.* 11, 52–65. doi: 10.1515/pjbr-2020-0005

Watson, D. P., Adams, E. L., Shue, S., Coates, H., McGuire, A., Chesher, J., et al. (2018). Defining the external implementation context: an integrative systematic literature review. *BMC Health Serv. Res.* 18:209. doi: 10.1186/s12913-018-3046-5

Zinina, A., Zaidelman, L., Arinkin, N., and Kotov, A. (2020). Non-verbal behavior of the robot companion: a contribution to the likeability. *Procedia Comput. Sci.* 169, 800–806. doi: 10.1016/j.procs.2020.02.160

TYPE Hypothesis and Theory
PUBLISHED 10 November 2022
DOI 10.3389/fpsyg.2022.949211

Check for updates

OPEN ACCESS

EDITED BY
Elizabeth A. Krusemark,
Millsaps College,
United States

REVIEWED BY
Ville J. Harjunen,
University of Helsinki,
Finland
Emanuel Jauk,
Medical University of Graz, Austria

*CORRESPONDENCE
Antonella Tramacere
a.tramacere@gmail.com

SPECIALTY SECTION
This article was submitted to Cognition,
a section of the journal
Frontiers in Psychology

RECEIVED 20 May 2022
ACCEPTED 20 October 2022
PUBLISHED 10 November 2022

CITATION
Tramacere A (2022) Face yourself: The social neuroscience of mirror gazing.
Front. Psychol. 13:949211.
doi: 10.3389/fpsyg.2022.949211

COPYRIGHT
© 2022 Tramacere. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Face yourself: The social neuroscience of mirror gazing

Antonella Tramacere[1,2]*

[1]Department of Philosophy and Communication Studies, University of Bologna, Bologna, Italy,
[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History (MPI-SHH), Jena, Germany

In philosophical and psychological accounts alike, it has been claimed that mirror gazing is like looking at ourselves *as* others. Social neuroscience and social psychology offer support for this view by showing that we use similar brain and cognitive mechanisms during perception of both others' and our own face. I analyse these premises to investigate the factors affecting the perception of one's own mirror image. I analyse mechanisms and processes involved in face perception, mimicry, and emotion recognition, and defend the following argument: because perception of others' face is affected by our feelings toward them, it is likely that feelings toward ourselves affect our responses to the mirror image. One implication is that negative self-feelings can affect mirror gazing instantiating a vicious cycle where the negative emotional response reflects a previously acquired attitude toward oneself. I conclude by discussing implications of this view for psychology and social studies.

KEYWORDS

mirror gazing, body image, body positivity, self-image, social neuroscience, social psychology, self-perception

Feelings in the mirror

When we perceive others' people in ecological situations, we may do a number of different things: we may mimic their facial expressions and resonate with their emotions, we may empathize or sympathize with them, or show appreciation or depreciation to them. Also, the way we feel toward others affects the way we perceive them and behavioral reactions towards them. For example, if we appreciate someone for their biography or personality, we will likely respond with positive emotions and prosocial behavior to their face (van Baaren et al., 2009; Franzen et al., 2018); on the contrary, if we do not like the other person, we may more probably lack to show sympathy and emotional connection to them.

What happens when we perceive our own self? What do we do, for example, when we look at our own face?

We cannot see our own face directly, but we can see it reflected in a mirror. Because of its autoscopic function, the mirror has fascinated human beings for centuries (Pendergrast, 2009). Through mirrors, we can perceive the visible aspects of our own face and body as others can see them and acquire an externalized perspective on ourselves. The mirror image is an objectified representation of ourselves and allow seeing us as through the gaze of an another.

Because mirror images embody an externalized perspective on the self, the ability to recognize oneself in the mirror has been considered the mark of a self-concept, namely a well-integrated, flexible, and conscious representation of the self as a being in the world (Gallup, 1977). The mirror test (Gallup, 1970, 1979), where experimenters place a dot on the forehead of the subject to see whether they try to touch or remove the dot, has been developed to inquire into animals and children ability to recognize themselves as themselves in the mirror, and to determine whether and when they become self-conscious.

In the last decades, this view has been challenged, and individuals who are able to recognize their body in the mirror are no longer considered to necessarily possess a self-concept (Heyes, 1994; Suddendorf and Butler, 2013). Further, variability has been found across human groups in mirror self-recognition, with different cultural groups showing different responses to the mirror test as consequence of their different practices with the mirror (Broesch et al., 2011). Not all cultures use the mirror for self-identification, and individuals with no experience with the mirror may interpret their mirror image differently (Rochat, 2009). For example, the Buryats of Eastern Mongolia conceive of mirrors as instruments that implement luminosity in the house and that enrich the display of precious objects; therefore, it is questionable whether individuals in these cultures have familiarized themselves with the mirror as a self-identifying tool (Humphrey, 2007).

Notwithstanding, in many societies the mirror is used as a tool to observe visible aspects of one's own body. In addition, the spread of self-directed pictures (aka selfie) in many countries across the globe suggests that, at least in these countries, most individuals recognize themselves as themselves in mirroring surfaces because of a process of socio-cultural learning (Rettberg, 2014). Consequently, for many individuals who are socialized with the mirror as a self-identifying tool, it makes sense to ask what factors affect the experience of mirror gazing (i.e., looking at ourselves in a mirroring surface).

Social psychology and neuroscience show that mechanisms of self-face perception are similar to mechanisms of others' face perception (Decety and Sommerville, 2003; Uddin et al., 2005; Bretas et al., 2021). This suggests that we perceive both ourselves and others by using common neurocognitive processes (brain and psychological mechanisms; Gallese, 2003). Neurocognitive findings about self and others' face perception are compatible with the idea that in the mirror we perceive the otherness of ourselves by adopting an external perspective on our own face; when we observe our own face in the mirror, we see it as it were the face of another. I will call this externalized perspective on one's own face "the social coding of mirror gazing." My contentious is that in the social coding of mirror gazing lay the keystone of the mirror as a tool of self-knowledge to inquire into how acquired feelings toward ourselves affect visual self-perception.

Findings in social psychology show that when we perceive others, our affective attitudes toward them modulate our responses to their face: consciously or unconsciously appreciating others affects whether, and to what extent, we respond with positive or negative emotions and corresponding facial mimicry (McIntosh, 2006; Bourgeois and Hess, 2008; van Baaren et al., 2009). Based on this premise, and on the similarity between the mechanisms of self and others' face perception, I argue that the perceptual processing of the mirror image is very likely influenced by the affective attitudes toward oneself. In other words, the way we feel toward ourselves affects the perception of ourselves in the mirror, and our behavioural and emotional responses to the mirror image.

The way we represent and evaluate our mirror image has important societal implications. It is widely claimed that a positive attitude toward oneself is connected to social wellbeing. For example, several social media and cultural movements advocate the importance of accepting and appreciating one's own self to acquire a positive *body image*, where the latter is typically defined as the perceptual and conceptual representation of, and the affective attitude toward one's own body (Sastre, 2014). Also, educators and psychologists in many different countries around the world acknowledge that *body positivity* is not a negligible feature of our psychological life, and it is pivotal for social well-being (Tylka and Wood-Barcalow, 2015).

Much work has been done to identify nature and origins of body image disturbance and eating disorders (Soh et al., 2006), to analyse possible psychological causes of these phenomena and social power dynamics involved in their emergence across groups and individuals (Sepúveda et al., 2002). However, to my knowledge less philosophical attention has been devoted to *whether* and *how* individuals' responses to the mirror image are affected by self-related feelings. One consequence of this is that the processes through which our feelings may have an impact on how we perceive and evaluate ourselves in the mirror are insufficiently investigated. My analysis aims to fill this gap.

In this paper, I investigate how affective attitudes may affect mirror self-perception through interpretations of the dynamics of top-down processes (from attitudes to perceptual responses) and the interplay between emotion and cognition. I take literally the hypothesis of a social coding of the mirror gazing, hence of an objectifying perspective on our mirror image, to studying phenomena of visual self-representation through social psychology. Consequently, perceiving ourselves as others becomes more than simply a metaphor describing our mirror experience, but rather an occasion to inquire into how what we think about ourselves affects self-perception.

The plan of the article is the following: in Section 2, I discuss results from social neuroscience showing that similar neurocognitive processes are activated during both self and others' face observation. In Section 3, I describe studies showing that others' face observation is affected by non-perceptual factors, such as affective attitudes toward others. In Section 4, I argue that mirror gazing is affected by the affective attitudes toward us. In Section 5 I discuss objections to this argument, and then conclude.

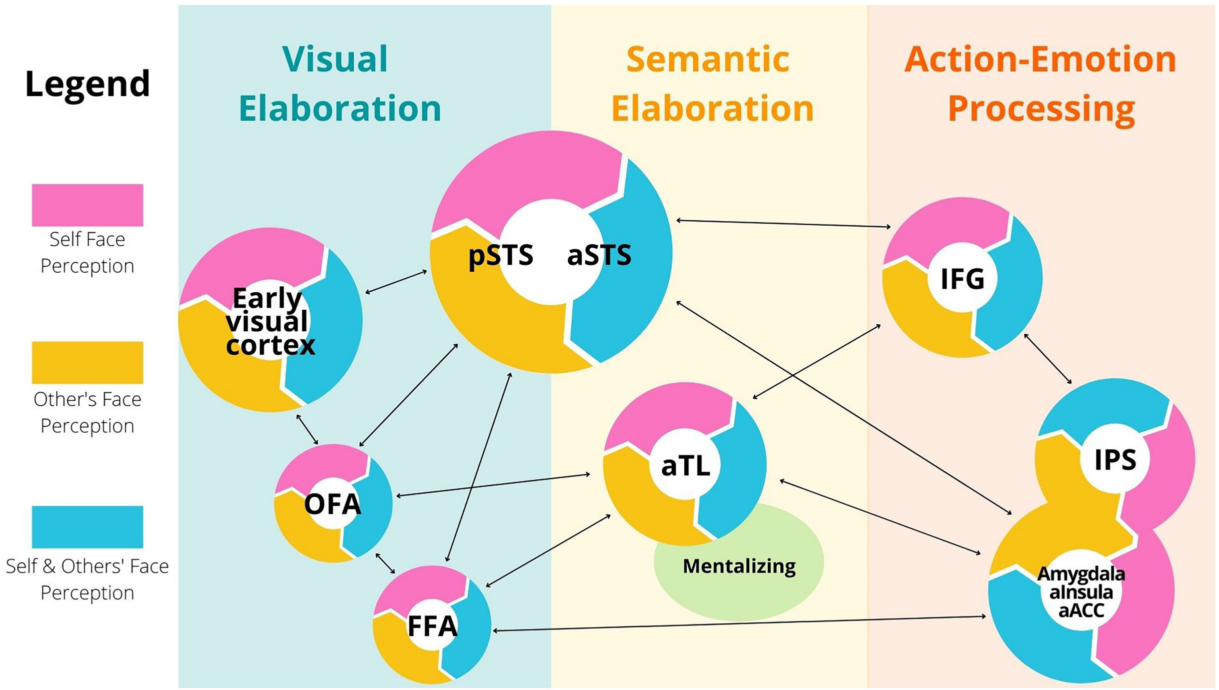


FIGURE 1
Simplified model of face selective areas. Different brain regions activate in a hierarchical and parallel fashion during perception of one's own face and during perception of others' faces. These regions are categorized by visual (light blue), semantic (light orange), action and emotion-centered elaboration (red). In this network, each region possesses both self-face perception areas (pink), others' face perception areas (yellow) and areas that activate in both conditions (blue). These distinctions are meant to reflect probabilistic activations during task performances, and not domain-specific functions of the brain. Abbreviations: OFA: Occipital Face Area, FFA: Face Fusiform Area, pSTS and aSTS: posterior and anterior Superior Temporal Sulcus respectively, aTL: anterior Temporal Lobule, IFG: Inferior Frontal Gyrus, IPS: Inferior Parietal Sulcus, aInsula: anterior Insula, ACC: Anterior Cingulate Cortex.

The brain in face perception

Similar brain mechanisms are activated during the observation of one's own face and of others' faces, where the similarity regards common brain location and neural circuitry. By discussing evidence from social neuroscience, I will show that neural regions with similar physiological properties activate during both self-face observation *and* other face observation (common coding) and that the regions of the brain that code for both self and others' faces are part of common brain networks (common neural circuitry).

I based the following discussion on the Haxby et al. (2000) model about the face processing network, its extension in the Duchaine and Yovel (2015) model and additional relevant studies. Through this body of research, we learned that specific areas of the primary visual cortex, plus regions of the occipital cortex, such as the Occipital Face Area (OFA), are active during the early visual elaboration of faces[1] (Figure 1, left blue panel). OFA is thought to code invariant structural features of faces, and to be involved in attribution of identity (Ambrus et al., 2017a,b). Additional important regions of the face processing network are the Fusiform Face Areas (FFA), and the anterior and posterior regions of the Superior Temporal Sulcus, respectively aSTS and pSTS [Figure 1, left (blue) and central (yellow) panels]. While the FFA is involved in holistic coding of faces, the anterior and posterior regions of STS multimodally code face expressions, through fine-grained processing of head, lip, and eye motion (McCarthy et al., 1997; Hoffman and Haxby, 2000). STS seems to be robustly involved in emotion recognition of perceived faces (Hoffman and Haxby, 2000; Engell and Haxby, 2007; Wang et al., 2016). In contrast, the function of FFA is still under debated, as it is yet unclear whether this area is involved in identity attribution, emotion recognition or both[2] (Bernstein and Yovel, 2015).

[1] Following Haxby et al. (2000) models, the Occipital Face Area is considered the gateway to the face processing network. However, there is evidence that multiple pathways convey face representations into the network (Dalrymple et al., 2011; Yang et al., 2016).

[2] Note that there is ongoing discussion on whether one common or two separate routes exist for face identity and face emotion processing. The majority of neuroimaging studies address both face identity and face expression functions, plus data are mixed. However, a detailed discussion of the ongoing controversies on the function of the various visual areas responding to faces is beyond the scope of the article.

Crucial for my argument here is that specific populations of neurons in the above mentioned face selective regions (OFA, FFA, aSTS and pSTS) are active during both self-face and others' face perception (Kircher et al., 2000; Platek et al., 2006; Figure 1, light blue sectors). There is no doubt that, in these brain areas, neurons also activate either during self-face or during others' face observation (Figure 1, yellow and pink sectors). In fact, although the areas typically active during perception of both self and other familiar faces are adjacent *and* partially overlapping, it is possible to experimentally disambiguate between them. Further, some studies found a right hemisphere dominance for self-face perception compared to others' face perception [even though other studies found equivalent bilateral activations for both self and others discrimination, probably reflecting differences in task context; see Platek et al. (2006) for a discussion of the latter point].

These findings are not surprising. We need to distinguish between different faces, and perceiving differences between self and others' faces also requires activating different brain and cognitive mechanisms. However, I do not think that these differences undermine the claim that brain regions coding for others' faces are similar to regions coding for self-face perception, because neurons active during both self and others' face observation are located in the same area, and they also share equivalent neurophysiological properties.

Sensorimotor and somatosensory neurons in frontal and parietal areas, as well as in the limbic system are also active during both self- and others' face perception (Uddin et al., 2005; Bretas et al., 2021). To support the idea that brain sensorimotor and somatosensory mechanisms for others' face perception are similar to those active for self-face perception, I will mainly discuss evidence regarding the mirror neuron system (MNS).

The MNS is a neural network including frontal and parietal areas [the frontoparietal circuits, comprising inferior frontal gyrus (IFG) and the inferior parietal sulcus (IPS)], but also limbic areas [anterior insula, anterior cingulate cortex (ACC) and the amygdala; Uddin et al., 2005; see Figure 1, red panel]. This wide cortical–subcortical network is called "mirror" because it contains many neurons with the key property of being active during both execution and perception of similar behaviour, including facial expressions (Ferrari et al., 2003, 2017). During face processing, wide neural populations in the temporal, limbic and frontoparietal circuits are active and code for fine-grained aspects of face actions and emotions, both in a social (allocentric) and individual (egocentric) condition (Uddin et al., 2005).

How do the mirror properties of frontoparietal and limbic regions of the brain support the claim that both self and others' face perception are coded by similar neurocognitive mechanisms? As I said, the MNS is a distributed network of neurons with mirroring properties, which are active both during observation and execution of self and others' face. Consider a smile. While you smile, a wide-spanning network of sensorimotor and somatosensory neurons activate and are associated with the time-course of your smile, coding both kinematics and valence information (Manjula et al., 2015). When you observe someone else smiling, a part of this network is also active and constitute an action-perception matching system for social cognition (Caruana et al., 2017). This basic matching mechanism underlie *perception* of disgust in self and others (Wicker et al., 2003), as well as pain (Timmers et al., 2018), laughter and joy (Caruana et al., 2017). Thus, perceiving others' facial expressions activates motor and somatosensory areas involved in the execution of the same facial behavior (Schilbach et al., 2008; Likowski et al., 2012).

Crucial for my argument is that sensorimotor and somatosensory neurons with mirror properties are also active while you observe your own face in the mirror. Studies have shown that the key areas of the frontoparietal network are activated during self-face movements and others' face perception, as well as during self-face perception (Decety and Sommerville, 2003). While watching their own face in the mirror, individuals activate portions of the frontoparietal MNS that are activated during others' faces perception (Platek et al., 2004, 2006; Uddin et al., 2005). Further, single neurons in the superior parietal cortex are active both during self-body observation in the mirror, during observations of others' body and during tactile perception of self-body (Bretas et al., 2021), suggesting that a similar coding mechanism might be present for face perception. Thus, beyond visual areas in the occipital and temporal cortices, also sensorimotor and somatosensory regions of the frontoparietal cortex activate during both self-face observation and others' face observation [Figure 1, right red panel].

The MNS is not the only associative circuit active both during self-face observation and others face observation. Selective areas of the anterior temporal lobule (aTL) and the mentalizing system, comprising dorsal prefrontal cortex, temporoparietal junction and anterior paracingulate cortex, are also active while we look at faces, in both an allocentric and egocentric perspective (Feinberg, 2001; Haxby et al., 2004; Morita et al., 2008; Platek et al., 2008; see Figure 1, yellow panel). During social cognitive tasks, regions in the MNS, the anterior temporal cortex, and the mentalizing systems work in conjunction. Cortical and subcortical neural nodes of the MNS are often involved in lower-order social cognitive mechanisms, such as identifying kinematic and affective aspects of observed behavior (Keysers et al., 2014; Urgesi et al., 2014; Carrillo et al., 2019). In contrast, the anterior temporal lobe and the nodes of the mentalizing system are thought to be involved in the semantic interpretation of others' goals, emotions and beliefs (Wong and Gallate, 2012; Hyatt et al., 2015).

The neurophysiological properties of the occipital and the temporal areas of the brain, the MNS and the mentalizing system active during various aspects of face perception support the view that both self and others' face observation are coded by common brain circuits. Again, as in the case of occipital and temporal cortices, also for the MNS and the mentalizing system, I talk of commonality and not of sameness, because neurons coding for one's own face and others' face are not identical nor 100% overlapping. I contend that the listed commonalities between brain mechanisms for self and others' face are sufficient for the generalization of functions from social cognition to mirror gazing.

Feelings in face-to-face interactions

In this section, I will analyze a series of studies supporting the view that the affective attitudes toward a person affects the perception of their face. Affective attitudes can be defined as feeling toward a person (or an object) and may include conscious and unconscious emotions for that person, and explicit evaluations about them such as appreciation for their biography or personality. As such, affective attitudes can be associated to many types of mental states, be propositional and non-propositional, conceptual and non-conceptual one[3]. Based on this definition, I will discuss studies that analyze how feelings toward a person affect perception of their face, even when those studies do not mention or define affective attitudes as I do here.

Affective attitudes toward others bias the perceptual process of their facial expressions, and this bias involves perceptual, emotional, and sensorimotor components. For example, previously acquired information about others affects neural processes in the occipitotemporal regions of the brain while perceiving their face (Abdel Rahman, 2011; Abdel Rahman and Sommer, 2012; Wieser et al., 2014). Knowing that someone is a rapist reduces activity in the visual STS during perception of their face, compared to when we think that they are kind people (Galli et al., 2006). These studies are yet inconclusive regarding the exact visual correlates of these changes. It is also unclear whether social information is processed by agents in rational or prerational terms, thus making it uncertain what level of mentalizing is required for the modulation of the perceptual process. All we know is that the differences in sensory regions of the observers correlate with the positive and negative recognition bias that they manifest during others' facial expressions.

When facial expressions are ambiguous, emotion recognition is biased along with priming of emotional descriptions, such as happy or sad (Halberstadt et al., 2009; Zhao et al., 2017). Biographic information with high emotional value about unknown individuals affects not only our eventual appreciation of them, but also the recognition of their facial expressions (Suess et al., 2015). In other words, affective knowledge about a person affects emotional responses to them and biases the recognition of their facial expressions toward the valence of the previously acquired information.

During the perception of others' faces, we seem to spontaneously retrieve information about the perceived person (Todorov et al., 2007). Interestingly, this information modulates our behavioral responses: individuals show a positive bias while perceiving facial expressions previously associated with positive personality features, by recognizing faster and more accurately happy faces than negative ones. At the same time, a negative bias is found with faces previously associated with negative personality features, because subjects are usually more accurate in categorizing negative expressions such as anger and sadness (Bijlstra et al., 2014; Albohn and Adams Jr., 2016). Interestingly, different behavioral responses during others' face perception are reflected in different patterns of brain activations (Abdel Rahman, 2011; Abdel Rahman and Sommer, 2012; Luo et al., 2016).

Individuals are more likely to manifest emotional contagion with people they like and feel emotionally connected to (Krebs, 1975). Emotional contagion occurs when an observer responds to others' emotional behavior with the same emotional expression (Zillman and Cantor, 1977; McIntosh, 2006). On the contrary, during the perception of strangers' faces, or of faces of people with whom there is no emotional connection, individuals show less emotional contagion (van Baaren et al., 2009). These responses have been detected quite robustly, and interestingly they often are conveyed bodily through changes in facial expressions, such as facial mimicry.

Facial mimicry occurs when individuals automatically react with same covert or overt facial movements to the facial behavior of others. Consider again a smile: smiling requires the joint activation of a series of facial muscles, controlling among other things the movements of the lip corner and of the ocular parts, such as the Orbicularis oculi and the Zygomaticus mayor (Manjula et al., 2015). When you perceive someone smiling, your Orbicularis oculi and Zygomaticus mayor muscles will also be activated. The story, as often is told in biology and psychology, is not so simple and linearly determined. Facial mimicry is modulated by a variety of factors (Bourgeois and Hess, 2008; Kraaijenvanger et al., 2017), and there is a bidirectional relationship between facial mimicry and social knowledge: Individuals mimicking more in response to others' facial expressions are normally rated as more likable and are more likely to trigger sympathy in the social partner (Duffy and Chartrand, 2015). In other words, responding with more facial muscles' activation during face-to-face interactions with others is likely going to make individuals nicer; at the same time, previously acquired sympathy or positive attitudes toward a person modulate the phenomenon of mimicry while perceiving their face (McIntosh, 2006; Bourgeois and Hess, 2008; Likowski et al., 2008; Kraaijenvanger et al., 2017).

When we observe the facial behavior of strangers or individuals we do not particularly like, we generally show less facial mimicry (Lakin and Chartrand, 2003). Consider one interesting and quite old study (McHugo et al., 1985), which

[3] I employ the philosophical concept of affective attitude to capture the entire spectrum of mental states that can be defined as feeling toward, hence states that regard the emotional salience of perceived objects and bias the perceptual process toward certain emotional modes. Employing psychological constructs such as "social knowledge" (e.g., knowledge about others) would instead force me to narrowly characterize the mental variables affecting perception of oneself in terms of conceptual content, while I want to remain open about the types of states that may influence the process of face perception. Because the way we feel toward someone or ourselves can be conveyed by both conceptual and non-conceptual content, I prefer to use the concept of affective attitudes in such a pluralistic way.

inquired into individuals' facial reactions to a Reagan's speech. The study found through electro magnetoencephalography that observers who did not support the U.S. President showed less activity in the cheek, and more corrugator brow activity than his supporters when viewing him smiling, suggesting that the perception of the smile of an enemy inhibits our mimicry response, and rather can bias us toward the expression of anger.

A series of studies confirms this trend, showing that people sympathy for the actor was correlating with the degree of facial mimicry showed (Chartrand and van Baaren, 2009; Duffy and Chartrand, 2015). The general pattern found was that more sympathy correlate with higher activation of cheek and mouth muscles involved in smiling and happy facial expression when the observed person was smiling and showing happiness. Similarly, during observed negative emotions in others, individuals were reacting with higher mimicry involving sad facial expressions. In contrast, when individuals have low sympathy for a person, they show less facial mimicry during perception of positive emotions, and increased tendency expression of negative emotions.

Interestingly, quite robust evidence suggests that the MNS is causally involved in phenomena of facial mimicry and emotional contagion (Hogeveen et al., 2015; Kraaijenvanger et al., 2017; Paz et al., 2022). This has been shown in the last decades through studies that have inquired simultaneously into the activity of the brain with more than one neuroscientific tool (Likowski et al., 2012). The simultaneous use of different neuroscience techniques with different direction of bias is often employed to disambiguate controversial results about causal questions (Tramacere, 2021). Although questions on the exact and functional role of the MNS remain, the claim that the MNS is causally involved in facial mimicry and related perception of others' emotion is relatively well established.

Note that evidence showing MNs activation during mimicry and emotional contagion does not imply that no other perception-motor neurons are active or important for explaining these phenomena. Further, the involvement of the MNS does say nothing on the functional model used to explain their effect and it is in principle compatible with different hypotheses (such as the simulation, direct perception and predictive coding hypothesis; Michael, 2011). The robust activation of the MNS during facial mimicry and emotional contagion episodes only says that the mechanisms of action-perception matching served by this system is important in explaining social phenomena based on face-to-face interactions (Tramacere and Ferrari, 2016).

Feeling toward the mirror image

The perception of our own face in the mirror may be affected by similar types of non-perceptual factors which modulate others' face perception in ecological situations. Specifically, the affective attitude toward ourselves can affect facial perceptual processing, as well as behavioural and psychological responses to our own mirror image. Therefore, affective self-attitudes have an impact on behavioral and psychological responses during self-face observation, and eventually what we know about the social brain can be instructive to inquire into the experience of mirror gazing.

During mirror gazing, individuals may activate regions of the social brain that convey responses in line with the internalized affective attitude toward themselves. The neurophysiological and circuitry properties of face perception network make plausible that as certain affective attitudes toward others bias us toward corresponding behavioural and emotional responses during their face perception, affective attitudes toward ourselves can bias behavioural and emotional responses to our own face in the mirror.

If my argument is correct, a negative way of representing oneself could produce negative emotions and corresponding facial expressions during mirror self-recognition. For example, an aversive self-image could (perhaps unconsciously) bias individuals' facial expressions toward certain emotions (sadness, disgust, and anger), and corresponding covert facial mimicry. Further, individuals with aversive self-image could show a higher activation of facial muscles normally activated during sadness, anger, or disgust, while an opposite bias could be found in subjects with positively connoted self-image. Note that while in the case of social cognitive responses, we could talk of emotional contagion and facial mimicry with the observed others, in the case of self-perception these concepts can only be used in a metaphorical way.

In the case of mirror gazing, negative behavioral and psychological responses toward oneself could be considered a *sui generis* form of emotional contagion, where the emotion that the subject resonates to relates to their own emotional attitude toward themselves. The emotional response to the mirror image may also trigger automatic and fast mimicry facial responses, so that the subject also covertly and unconsciously activates facial muscles that correspond to negative emotional responses, such as contempt, disgust, anger, or sadness. In other words, a self-sustaining vicious circle could be instantiated during mirror gazing, involving various forms of negative responses toward oneself.

As far as I am aware, no studies tested this specific hypothesis. However, various psychological studies are compatible with and provide broad support for my claim. For example, an extensive range of studies have showed that body concerns and self-esteem are bidirectionally related (Feingold, 1992; O'Dea, 2012; Felisberti, 2014). In one study (Oikawa et al., 2012) people reported low self-appreciation and low self-esteem when their image was compared with individuals who were rated as more attractive, suggesting that the affective attitudes toward oneself is not fix across time, but dynamical and influenced by contextual factors. In this study, self-face appreciation was associated with activation of the reward system, while negative self-evaluation modulated areas of the face perception network, supporting the view that a positive self-image produces positive feelings.

One study (Jauk et al., 2017) inquired into whether personality disorders can affect self-face evaluation, and the findings suggest that this might be the case. The authors showed that, compared to typical subjects, subjects with high scores of narcissism have greater activation in areas of the brain which are typically correlated with expectancy violation and negative emotion. Another study (Potthoff and Schienle, 2021) performed with eye-tracking show that subjects with low self-esteem look longer at their own face, possibly reflecting a higher critical gaze on oneself.

Further studies have found correlations between affective self-knowledge and psychological and behavioral responses during perception of one's own face and body. One study tested emotional responses to distorted self-face perception, where subjects rated altered images of their own face as more embarrassing than the altered image of others. Interestingly, while recognition of self-face correlated with the activity of the action MNS, changes in embarrassment were co-varying with activity of both the MNS and the mentalizing system (Morita et al., 2008). Another recent study (Maister et al., 2021) inquired into the pictorial visual representation of individuals and compared it with various self-construal index, and showed that the valence of individuals self-representation correlated with self-attributed visual features.

A shortcoming of these studies is that the causal direction of interaction is not inquired about; therefore, other causal factors could produce self-directed emotions with, e.g., negative valence, and the emotional attitude toward oneself could be a consequence, and not a cause of those results. However, the validity of my argument does not require that no other factors can affect self-face perception and associated behavioural and psychological responses, nor it requires that self-perceived physical and psychological characteristics are univocally, rather than bidirectionally, related. I will engage with objections to my argument in the next section.

Objections

There could be objections to the argument that feelings toward oneself affect the perception of one's own face and corresponding behavioral and psychological responses. I will consider two main objections: (1) causal effect from non-perceptual processes (such as affective attitudes) to mirror gazing are unlikely, because the emotional ways we represent objects cannot exert influence on perceptual responses; (2) even if (somehow) responses to others' face could be influenced by affective attitudes toward others, this process cannot generalize to perception of oneself.

I will address these objections in turn, beginning with (1). Someone could object that although some scholars claim that non-perceptual content affects perception (Stokes, 2013), this claim is still controversial and is not supported by conclusive arguments nor evidence. Furthermore, the objection would continue, none of the studies that I have discussed here conclusively show that non-perceptual content, such as attitudes, have a direct causal and semantic influence on perceptual responses, such as visual perception and object representation. Therefore, according to this objection, the conclusion that affective attitudes influence responses during observation of one' own face in the mirror is wrong or at least unsupported.

I think this objection is out of target. It is true that, in my analysis, I consider modulations from non-perceptual processes, such as affective attitudes, to perceptual mechanisms. However, I am not claiming that this modulation regards low-level, basic visual properties of the observed object, such as the invariant aspects of face identity perception. I am not analyzing the impact of higher-level content (beliefs, intentions, and desires) on low-level visual coding of face. My argument is tangential to cognitive penetration debate, which regards whether cognition affects perceptual processes, with perception being narrowly defined as a purely sensory, non-interpretative process, and cognition being defined as elaboration in propositional and conceptual terms.

My focus is rather mostly on higher-order processes of perception, where the multimodal sensory coding of a percept (i.e., faces) overlaps and intermingles with motor and affective coding. I have thus embraced here a rich conception of perception (Burnston, 2017, in press) and inquired into how these higher-level features of perception (seeing as) are connected to behavioral and psychological responses of an individual (e.g., emotional responses and facial mimicry). If you agree that perceptual processes do not necessarily occur in encapsulated and domain-specific areas of the brain which are insensitive to processes in other areas and domains, and if you allow perceptual responses to involve and recruit emotional, affective, and motor responses, the objection that attitudes cannot affect perception will lose force.

Recall previously discussed evidence. We have seen that during observations of others' face, brain processes in the occipitotemporal cortices are modulated by the ways we represent others. We have seen however that it is unclear what are the functional correlates of those changes. Similarly, it is likely that during the observation of self-face, patterns of changes in the facial perceptual stream in the occipitotemporal lobe correlate with the valence of attitude toward oneself, but I do not think we can make any reasonable prediction about what these changes are in the subjects' eyes.

Further, during observation of others' face, brain changes in somatosensory and sensorimotor areas predict patterns of facial mimicry and emotional responses to others' people face, and these responses are modulated by affective attitudes toward others. Based on the arguments I offered in previous sections, it is reasonable that having aversive self-image produce changes in the face processing network, and that these changes bias the activation of negative expressions, such as sadness. Note however that even though this prediction will be verified, and that we can find a correlation between negative self-image and sadness during mirror perception, I do not think that we can be sure that the subject of this experience is feeling sad. It is possible that a subject showing brain activation, bodily markers and facial mimicry responses normally correlated with sadness also feels sad, but this is not obvious. Like in the case of others' face perception, during perception of one's own face the emotional response to oneself could remain unconscious, thus making the mental state attribution to the subject arduous. Although we cannot identify the exact mental correlates of subjects' behavioral and emotional responses during mirror gazing, we can base our analyses on the probabilistic relations between facial expression and associated emotions to narrow the space of inferences about subjects' experiences with the mirror.

Let us see objection (2). One could claim that even if affective attitudes toward others may influence responses of the observer during others' face perception, this does not allow generalizing this premises to self-face observation. On this objection, although the brain mechanisms for self and others' face observations are similar, (i) their activation says little on the similarity between self and others' face representation at the psychological level. Further, (ii) similar brain mechanisms for self and others' face perception do not ensure that they are affected by similar psychological mechanisms, such as, feeling toward oneself. Affective attitudes toward oneself and others may be instantiated by different mechanisms, and we do not know whether they can modulate subjects' responses to one's own face, as they modulate responses to others' face.

Regarding (i), note that while I have examined the similarities between brain mechanisms of self- and other's face perception in an analytical way, a broad range of studies already provide support that self and others' face perception share many similar aspects at the psychological levels; these studies suggest that we recognize others' faces through psychological mechanisms that are similar to those while discriminating our own faces (Rochat, 2009; Rochat et al., 2012; Porciello et al., 2018). To contradict this claim, one should demonstrate that the perception of our own face has distinctive psychological features, and that these features prevent our own face perception to be modulated by emotions and feelings about ourselves. I honestly do not see any evidence pointing in this direction, and already existing studies support the conclusion that this is not the case [see for example Oikawa et al. (2012)].

The objection (ii) puts doubt that responding with, say, sadness to one's own face during self-face observation is caused by generalized negative feelings toward oneself, rather than by alternative causes. The objection could add that it is possible individuals show dissociations between fast emotional/mimicry responses on the one hand, and explicit, reportable emotional attitudes toward oneself on the other hand. That is, if we observe bodily markers, brain activations and facial mimicry patterns that are normally correlated with sadness during mirror gazing, not only we cannot conclude that the subject is feeling sad, but we cannot even say that this response is caused by an aversive self-image. Sadness or other negatively connoted emotional responses to one's own face could be caused by other psychological or non-psychological mechanisms, and not necessarily by aversive self-affective attitudes.

I agree that multiple causes can be responsible for aversive responses during mirror gazing. Further, I acknowledge the complexity and possible dissociation between automatic emotional responses and more reflective evaluative representation of oneself. I do not think however that this complexity speaks against the validity (and heuristic utility) of my argument. Several methods could be employed to make sure that the sad response of the subject is elicited by self-representation during mirror gazing, for example by priming subjects' responses to others' faces with valence information about oneself [one study in this direction is again Oikawa et al. (2012)].

While many methods can be employed for narrowing down the inferences that self-related affective attitude can bias responses to one's own face, it is not the purpose of this paper to propose exactly which experimental methods can settle the debate. My interest here is only to provide good enough reasons for possible neurocognitive mechanisms that can explain whether and how attitudes toward oneself affect mirror gazing. This explanation can do further justice to psychological data that we already possess, and that can possibly describe social and cultural phenomena involving affective self-representation and perception of oneself in the mirror.

Conclusion

For decades cognitive neuroscience has told us that we use parts of the brain involved in performing actions and emotions to perceive and understand actions and emotions of others. We know others through the neurocognitive structures that we use for moving, sensing, and feeling in the world (Cacioppo, 2006); but we also sense, feel, and get to know ourselves through the neurocognitive structures that we have acquired during affective and communicative interactions with others. Because we are not able to directly perceive our face through vision and the first knowledge that we acquired about faces derives from facial interactions with others, the reciprocity of self-other perception is especially relevant for self-face observation on mirroring devices.

In this paper, I have inquired about self-face perception through the lenses of social psychology and neuroscience. Analysing mirror gazing through social neuroscience does not aim at reducing the phenomenon of face self-perception to the activation of parts of the brain active during the visual perception of our own face. I instead considered neural and behavioral evidence as an occasion to enrich our understanding of the experience of mirror gazing and stimulate new thinking in the philosophy of mind and experimental psychology.

A social neuroscience approach to mirror gazing is centered on addressing what happens while we see our own face in mirroring devices, what responses we show in front of the mirror image, and whether these responses may say something about the way we represent ourselves. I have discussed studies supporting the view that similar neurocognitive mechanisms, in respect to both brain location and neural circuitry, are active for both self-face and others' face perception. Specific activation in key visual areas, the action and emotion MNS and the mentalizing system suggest that during self-face observation, the neurocognitive mechanisms involved in the perceptual, action, and emotion coding of others' face are modulated by non-perceptual variables, such as affective attitudes toward oneself. The reviewed studies suggest that self-affective attitudes could affect whether we respond with positive or negative emotions to oneself, and with a corresponding facial mimicry, and these responses could be mediated by the face processing areas, the action and emotion MNSs and the mentalizing system.

If the way we feel toward ourselves can produce a series of negative or positive emotional responses to the mirror image, they may trigger a vicious circle which involves various forms of antipathy and depreciation toward oneself. Since these responses are likely to be automatic and fast, this negative circle could be difficult to break. Therefore, the hypotheses I formulated imply that face-to-face interactions are relevant to the appreciation, and understanding of ourselves, and that the interactions with others are relevant to delineate the phenomenological experience with oneself. Mirror gazing would then necessarily be included in the boundary of social interactions, and how the gaze of others affects the perception and understanding of oneself.

I am convinced that the arguments proposed in this article can be informative for psychological and phenomenological research, by providing a heuristic parallel between social and self-related cognitive processes, which can explain important societal phenomena and pave the way to novel experimental predictions to be explored in future research. If my interpretations are correct, my arguments of a bias from affective attitude to self-face perception can provide an important basis for making sense of the rich experience of mirror gazing in many contemporary cultures.

According to some psychological studies (e.g., Perugi et al., 1997), individuals who show negative attitudes toward one's own face or body are not necessarily considered ugly or unpleasant by others. In some cases, the perceptual distortion during self-observation can be so prominent to produce psychological disorders, such as dysmorphophobia, namely a psychological condition characterized by the phobia of being ugly and by the pathological use of mirrors, which can produce significant discomfort in the individuals that are affected by it (Veale and Riley, 2001). Because often no significant correlations have been found between objective bodily features (as evaluated by other individuals) and body-image, the hypothesis of a role of affective self-attitude in self-perception can be relevant to explain such phenomena. I contend that these hypotheses can provide an interesting basis for analysing the phenomenology of mirror gazing in individuals of different age and developmental history, and of the mirror as a tool for self-knowledge.

Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Author contributions

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

Acknowledgments

I would like to thank Marco Viola and Francesco Ellia for critical reading and suggestions on the paper, Marzio Gerbella and Luigi Lobaccaro for helping me with, respectively, the content and design of the Figure 1.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdel Rahman, R. (2011). Facing good and evil: early brain signatures of affective biographical knowledge in face recognition. *Emotion* 11, 1397–1405. doi: 10.1037/a0024717

Abdel Rahman, R., and Sommer, W. (2012). Knowledge scale effects in face recognition: an electrophysiological investigation. *Cogn. Affect. Behav. Neurosci.* 12, 161–174. doi: 10.3758/s13415-011-0063-9

Albohn, D. N., and Adams, R. B. Jr. (2016). "Social vision: at the intersection of vision and person perception" in *Neuroimaging Personality, Social Cognition, and Character*. eds. J. R. Absher and J. Cloutier (Netherlands: Elsevier Academic Press), 159–186.

Ambrus, G. G., Dotzer, M., Schweinberger, S. R., and Kovács, G. (2017a). The occipital face area is causally involved in the formation of identity-specific face representations. *Brain Struct. Funct.* 222, 4271–4282. doi: 10.1007/s00429-017-1467-2

Ambrus, G. G., Windel, F., Burton, A. M., and Kovács, G. (2017b). Causal evidence of the involvement of the right occipital face area in face-identity acquisition. *NeuroImage* 148, 212–218. doi: 10.1016/j.neuroimage.2017.01.043

Bernstein, M., and Yovel, G. (2015). Two neural pathways of face processing: a critical evaluation of current models. *Neurosci. Biobehav. Rev.* 55, 536–546. doi: 10.1016/j.neubiorev.2015.06.010

Bijlstra, G., Holland, R. W., Dotsch, R., Hugenberg, K., and Wigboldus, D. H. J. (2014). Stereotype associations and emotion recognition. *Personal. Soc. Psychol. Bull.* 40, 567–577. doi: 10.1177/0146167213520458

Bourgeois, P., and Hess, U. (2008). The impact of social context on mimicry. *Biol. Psychol.* 77, 343–352. doi: 10.1016/j.biopsycho.2007.11.008

Bretas, R., Taoka, M., Hihara, S., Cleeremans, A., and Iriki, A. (2021). Neural evidence of Mirror self-recognition in the secondary somatosensory cortex of macaque: observations from a single-cell recording experiment and implications for consciousness. *Brain Sci.* 11:157. doi: 10.3390/brainsci11020157

Broesch, T., Callaghan, T., Henrich, J., Murphy, C., and Rochat, P. (2011). Cultural variations in Children's Mirror self-recognition. *J. Cross-Cult. Psychol.* 42, 1018–1029. doi: 10.1177/0022022110381114

Burnston, D. C. (2017). Is aesthetic experience evidence for cognitive penetration? *New Ideas Psychol.* 47, 145–156. doi: 10.1016/j.newideapsych.2017.03.012

Burnston, D. (in press). *Perceptual Learning, Categorical Perception, and Cognitive Permeation.* Dialectica.

Cacioppo, J. T. (2006). Social Neuroscience. *Am. J. Psychol.* 119, 664–668. doi: 10.2307/20445370

Carrillo, M., Han, Y., Migliorati, F., Liu, M., Gazzola, V., and Keysers, C. (2019). Emotional Mirror neurons in the Rat's anterior cingulate cortex. *Curr. Biol.* 29, 1301–1312. doi: 10.1016/j.cub.2019.03.024

Caruana, F., Avanzini, P., Gozzo, F., Pelliccia, V., Casaceli, G., and Rizzolatti, G. (2017). A mirror mechanism for smiling in the anterior cingulate cortex. *Emotion* 17, 187–190. doi: 10.1037/emo0000237

Chartrand, T. L., and van Baaren, R. (2009). "Chapter 5 human mimicry," in *Advances in Experimental Social Psychology*, *Vol. 41* (Cambridge: Academic Press), 219–274.

Dalrymple, K. A., Oruç, I., Duchaine, B., Pancaroglu, R., Fox, C. J., Iaria, G., et al. (2011). The anatomic basis of the right face-selective N170 IN acquired prosopagnosia: a combined ERP/fMRI study. *Neuropsychologia* 49, 2553–2563. doi: 10.1016/j.neuropsychologia.2011.05.003

Decety, J., and Sommerville, J. A. (2003). Shared representations between self and other: a social cognitive neuroscience view. *Trends Cogn. Sci.* 7, 527–533. doi: 10.1016/j.tics.2003.10.004

Duchaine, B., and Yovel, G. (2015). A revised neural framework for face processing. *Annual Review of Vision Science* 1, 393–416. doi: 10.1146/annurev-vision-082114-035518

Duffy, K. A., and Chartrand, T. L. (2015). Mimicry: causes and consequences. *Curr. Opin. Behav. Sci.* 3, 112–116. doi: 10.1016/j.cobeha.2015.03.002

Engell, A. D., and Haxby, J. V. (2007). Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia* 45, 3234–3241. doi: 10.1016/j.neuropsychologia.2007.06.022

Feinberg, T. E. (2001). *Altered Egos: How the Brain Creates the Self.* Oxford: Oxford university press.

Feingold, A. (1992). Good-looking people are not what we think. *Psychol. Bull.* 111, 304–341. doi: 10.1037/0033-2909.111.2.304

Felisberti, F. M. K. (2014). Self-face perception: individual differences and discrepancies associated with mental self-face representation, attractiveness and self-esteem. *Psychol. Neurosci.* 7, 65–72. doi: 10.3922/j.psns.2014.013

Ferrari, P. F., Gallese, V., Rizzolatti, G., and Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *Eur. J. Neurosci.* 17, 1703–1714. doi: 10.1046/j.1460-9568.2003.02601.x

Ferrari, P. F., Gerbella, M., Coudé, G., and Rozzi, S. (2017). Two different mirror neuron networks: the sensorimotor (hand) and limbic (face) pathways. *Neuroscience* 358, 300–315. doi: 10.1016/j.neuroscience.2017.06.052

Franzen, A., Mader, S., and Winter, F. (2018). Contagious yawning, empathy, and their relation to prosocial behavior. *J. Exp. Psychol. Gen.* 147, 1950–1958. doi: 10.1037/xge0000422

Gallese, V. (2003). The roots of empathy: the shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology* 36, 171–180. doi: 10.1159/000072786

Galli, G., Feurra, M., and Viggiano, M. P. (2006). "Did you see him in the newspaper?" electrophysiological correlates of context and valence in face processing. *Brain Res.* 1119, 190–202. doi: 10.1016/j.brainres.2006.08.076

Gallup, G. G. (1970). Chimpanzees: self-recognition. *Science* 167, 86–87. doi: 10.1126/science.167.3914.86

Gallup, G. G. (1977). Self recognition in primates: a comparative approach to the bidirectional properties of consciousness. *Am. Psychol.* 32, 329–338. doi: 10.1037/0003-066X.32.5.329

Gallup, G. G. (1979). "Self-recognition in chimpanzees and man: a developmental and comparative perspective," in *Genesis of behavior*. eds. M. Lewis and M. Rosenblum, *Vol. 2* (New York: Plenum Press), 107–126.

Halberstadt, J., Winkielman, P., Niedenthal, P. M., and Dalle, N. (2009). Emotional conception: how embodied emotion concepts guide perception and facial action. *Psychol. Sci.* 20, 1254–1261. doi: 10.1111/j.1467-9280.2009.02432.x

Haxby, J. V., Gobbini, M. I., and Montgomery, K. (2004). "Spatial and temporal distribution of face and object representations in the human brain," in *The cognitive neurosciences. 3rd Edn.* ed. M. S. Gazzaniga (Cambridge: Boston Review), 889–904.

Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends Cogn. Sci.* 4, 223–233. doi: 10.1016/S1364-6613(00)01482-0

Heyes, C. M. (1994). Reflections on self-recognition in primates. *Anim. Behav.* 47, 909–919. doi: 10.1006/anbe.1994.1123

Hoffman, E. A., and Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nat. Neurosci.* 3, 80–84. doi: 10.1038/71152

Hogeveen, J., Chartrand, T. L., and Obhi, S. S. (2015). "Social mimicry enhances mu-suppression during action observation" in *Cerebral Cortex*, *Vol. 25* (New York, N.Y: Oxford University Press), 2076–2082.

Humphrey, C. (2007). Inside and outside the Mirror: Mongolian shamans' mirrors as instruments of Perspectivism. *Inner Asia* 9, 173–195. doi: 10.1163/146481707793646557

Hyatt, C. J., Calhoun, V. D., Pearlson, G. D., and Assaf, M. (2015). Specific default mode subnetworks support mentalizing as revealed through opposing network recruitment by social and semantic FMRI tasks. *Hum. Brain Mapp.* 36, 3047–3063. doi: 10.1002/hbm.22827

Jauk, E., Benedek, M., Koschutnig, K., Kedia, G., and Neubauer, A. C. (2017). Self-viewing is associated with negative affect rather than reward in highly narcissistic men: an fMRI study. *Sci. Rep.* 7:1. doi: 10.1038/s41598-017-03935-y

Keysers, C., Perrett, D. I., and Gazzola, V. (2014). Hebbian learning is about contingency, not contiguity, and explains the emergence of predictive mirror neurons. *Behav. Brain Sci.* 37, 205–206. doi: 10.1017/S0140525X13002343

Kircher, T. T. J., Senior, C., Phillips, M. L., Benson, P. J., Bullmore, E. T., Brammer, M., et al. (2000). Towards a functional neuroanatomy of self processing: effects of faces and words. *Cogn. Brain Res.* 10, 133–144. doi: 10.1016/S0926-6410(00)00036-7

Kraaijenvanger, E. J., Hofman, D., and Bos, P. A. (2017). A neuroendocrine account of facial mimicry and its dynamic modulation. *Neurosci. Biobehav. Rev.* 77, 98–106. doi: 10.1016/j.neubiorev.2017.03.006

Krebs, D. (1975). Empathy and altruism. *J. Pers. Soc. Psychol.* 32, 1134–1146. doi: 10.1037/0022-3514.32.6.1134

Lakin, J. L., and Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychol. Sci.* 14, 334–339. doi: 10.1111/1467-9280.14481

Likowski, K. U., Mühlberger, A., Gerdes, A. B. M., Wieser, M. J., Pauli, P., and Weyers, P. (2012). Facial mimicry and the mirror neuron system: simultaneous acquisition of facial electromyography and functional magnetic resonance imaging. *Front. Hum. Neurosci.* 6:214. doi: 10.3389/fnhum.2012.00214

Likowski, K. U., Mühlberger, A., Seibt, B., Pauli, P., and Weyers, P. (2008). Modulation of facial mimicry by attitudes. *J. Exp. Soc. Psychol.* 44, 1065–1072. doi: 10.1016/j.jesp.2007.10.007

Luo, Q. L., Wang, H. L., Dzhelyova, M., Huang, P., and Mo, L. (2016). Effect of affective personality information on face processing: evidence from ERPs. *Front. Psychol.* 7:810. doi: 10.3389/fpsyg.2016.00810

Maister, L., De Beukelaer, S., Longo, M., and Tsakiris, M. (2021). The self in the mind's eye: revealing how we truly see ourselves through reverse correlation. *Psychol. Sci.* 32, 1965–1978. doi: 10.1177/09567976211018618

Manjula, W. S., Sukumar, M. R., Kishorekumar, S., Gnanashanmugam, K., and Mahalakshmi, K. (2015). Smile: a review. *J. Pharm. Bioallied Sci.* 7, 273–275. doi: 10.4103/0975-7406.155951

McCarthy, G., Puce, A., Gore, J. C., and Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *J. Cogn. Neurosci.* 9, 605–610. doi: 10.1162/jocn.1997.9.5.605

McHugo, G. J., Lanzetta, J. T., Sullivan, D. G., Masters, R. D., and Englis, B. G. (1985). Emotional reactions to a political leader's expressive displays. *J. Pers. Soc. Psychol.* 49, 1513–1529. doi: 10.1037/0022-3514.49.6.1513

McIntosh, D. N. (2006). Spontaneous facial mimicry, liking and emotional contagion. *Pol. Psychol. Bull.* 37, 31–42.

Michael, J. (2011). Four models of the functional contribution of mirror systems. *Philos. Explor.* 14, 185–194. doi: 10.1080/13869795.2011.569747

Morita, T., Itakura, S., Saito, D. N., Nakashita, S., Harada, T., Kochiyama, T., et al. (2008). The role of the right prefrontal cortex in self-evaluation of the face: a functional magnetic resonance imaging study. *J. Cogn. Neurosci.* 20, 342–355. doi: 10.1162/jocn.2008.20024

O'Dea, J. A. (2012). "Body image and self-esteem," in *Encyclopedia of Body Image and Human Appearance*, *Vol. 1* (Netherlands: Elsevier Academic Press), 141–147.

Oikawa, H., Sugiura, M., Sekiguchi, A., Tsukiura, T., Miyauchi, C. M., Hashimoto, T., et al. (2012). Self-face evaluation and self-esteem in young females:

an fMRI study using contrast effect. *NeuroImage* 59, 3668–3676. doi: 10.1016/j.neuroimage.2011.10.098

Paz, L. V., Viola, T. W., Milanesi, B. B., Sulzbach, J. H., Mestriner, R. G., Wieck, A., et al. (2022). Contagious depression: automatic mimicry and the mirror neuron system - a review. *Neurosci. Biobehav. Rev.* 134:104509. doi: 10.1016/j.neubiorev.2021.12.032

Pendergrast, M. (2009). *Mirror, Mirror: A History of the Human Love Affair With Reflection*. Hachette: United Kingdom.

Perugi, G., Giannotti, D., Frare, F., Vaio, S. D., Valori, E., Maggi, L., et al. (1997). Prevalence, phenomenology and comorbidity of body dysmorphic disorder (dysmorphophobia) in a clinical population. *Int. J. Psychiatry Clin. Pract.* 1, 77–82. doi: 10.3109/13651509709024707

Platek, S. M., Keenan, J. P., Gallup, G. G., and Mohamed, F. B. (2004). Where am I? The neurological correlates of self and other. *Brain Res. Cogn. Brain Res.* 19, 114–122. doi: 10.1016/j.cogbrainres.2003.11.014

Platek, S. M., Loughead, J. W., Gur, R. C., Busch, S., Ruparel, K., Phend, N., et al. (2006). Neural substrates for functionally discriminating self-face from personally familiar faces. *Hum. Brain Mapp.* 27, 91–98. doi: 10.1002/hbm.20168

Platek, S. M., Wathne, K., Tierney, N. G., and Thomson, J. W. (2008). Neural correlates of self-face recognition: an effect-location meta-analysis. *Brain Res.* 1232, 173–184. doi: 10.1016/j.brainres.2008.07.010

Porciello, G., Bufalari, I., Minio-Paluello, I., Di Pace, E., and Aglioti, S. M. (2018). The 'enfacement' illusion: a window on the plasticity of the self. *Cortex* 104, 261–275. doi: 10.1016/j.cortex.2018.01.007

Potthoff, J., and Schienle, A. (2021). Effects of self-esteem on self-viewing: an eye-tracking investigation on Mirror gazing. *Behavioral Sciences* 11:164. doi: 10.3390/bs11120164

Rettberg, J. W. (2014). *Seeing ourselves through technology: How we use Selfies, Blogs and Wearable Devices to See and Shape Ourselves*. Germany: Springer.

Rochat, P. (2009). *Others in mind: Social origins of self-consciousness* (pp. 10–253). Cambridge: Cambridge University Press.

Rochat, P., Broesch, T., and Jayne, K. (2012). Social awareness and early self-recognition. *Conscious. Cogn.* 21, 1491–1497. doi: 10.1016/j.concog.2012.04.007

Sastre, A. (2014). Towards a radical body positive: Ingenta. *Connect* 14, 929–943. doi: 10.1080/14680777.2014.883420

Schilbach, L., Eickhoff, S. B., Mojzisch, A., and Vogeley, K. (2008). What's in a smile? Neural correlates of facial embodiment during social interaction. *Soc. Neurosci.* 3, 37–50. doi: 10.1080/17470910701563228

Sepúveda, A. R., Botella, J., and León, J. A. (2002). Body-image disturbance in eating disorders: a meta-analysis. *Psychology in Spain* 6, 83–95.

Soh, N. L., Touyz, S. W., and Surgenor, L. J. (2006). Eating and body image disturbances across cultures: a review. *Eur. Eat. Disord. Rev.* 14, 54–65. doi: 10.1002/erv.678

Stokes, D. (2013). Cognitive penetrability of perception. *Philosophy Compass* 8, 646–663. doi: 10.1111/phc3.12043

Suddendorf, T., and Butler, D. L. (2013). The nature of visual self-recognition. *Trends Cogn. Sci.* 17, 121–127. doi: 10.1016/j.tics.2013.01.004

Suess, F., Rabovsky, M., and Abdel Rahman, R. (2015). Perceiving emotions in neutral faces: expression processing is biased by affective person knowledge. *Soc. Cogn. Affect. Neurosci.* 10, 531–536. doi: 10.1093/scan/nsu088

Timmers, I., Park, A. L., Fischer, M. D., Kronman, C. A., Heathcote, L. C., Hernandez, J. M., et al. (2018). Is empathy for pain unique in its neural correlates? A meta-analysis of neuroimaging studies of empathy. *Front. Behav. Neurosci.* 12:289. doi: 10.3389/fnbeh.2018.00289

Todorov, A., Gobbini, M. I., Evans, K. K., and Haxby, J. V. (2007). Spontaneous retrieval of affective person knowledge in face perception. *Neuropsychologia* 45, 163–173. doi: 10.1016/j.neuropsychologia.2006.04.018

Tramacere, A. (2021). *Triangulating tools in the messiness of cognitive neuroscience. In The Tools of Neuroscience Experiment*. London. Routledge.

Tramacere, A., and Ferrari, P. F. (2016). Faces in the mirror, from the neuroscience of mimicry to the emergence of mentalizing. *Journal of Anthropological Sciences = Rivista Di Antropologia: JASS* 94, 113–126. doi: 10.4436/JASS.94037

Tylka, T. L., and Wood-Barcalow, N. L. (2015). What is and what is not positive body image? Conceptual foundations and construct definition. *Body Image* 14, 118–129. doi: 10.1016/j.bodyim.2015.04.001

Uddin, L. Q., Kaplan, J. T., Molnar-Szakacs, I., Zaidel, E., and Iacoboni, M. (2005). Self-face recognition activates a frontoparietal "mirror" network in the right hemisphere: an event-related fMRI study. *NeuroImage* 25, 926–935. doi: 10.1016/j.neuroimage.2004.12.018

Urgesi, C., Candidi, M., and Avenanti, A. (2014). Neuroanatomical substrates of action perception and understanding: an anatomic likelihood estimation meta-analysis of lesion-symptom mapping studies in brain injured patients. *Front. Hum. Neurosci.* 8:344. doi: 10.3389/fnhum.2014.00344

van Baaren, R., Janssen, L., Chartrand, T. L., and Dijksterhuis, A. (2009). Where is the love? The social aspects of mimicry. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 2381–2389. doi: 10.1098/rstb.2009.0057

Veale, D., and Riley, S. (2001). Mirror, mirror on the wall, who is the ugliest of them all? The psychopathology of mirror gazing in body dysmorphic disorder. *Behav. Res. Ther.* 39, 1381–1393. doi: 10.1016/S0005-7967(00)00102-9

Wang, X., Song, Y., Zhen, Z., and Liu, J. (2016). Functional integration of the posterior superior temporal sulcus correlates with facial expression recognition. *Hum. Brain Mapp.* 37, 1930–1940. doi: 10.1002/hbm.23145

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., and Rizzolatti, G. (2003). Both of us disgusted in my insula: the common neural basis of seeing and feeling disgust. *Neuron* 40, 655–664. doi: 10.1016/S0896-6273(03)00679-2

Wieser, M. J., Gerdes, A. B. M., Büngel, I., Schwarz, K. A., Mühlberger, A., and Pauli, P. (2014). Not so harmless anymore: How context impacts the perception and electrocortical processing of neutral faces. *NeuroImage* 92, 74–82. doi: 10.1016/j.neuroimage.2014.01.022

Wong, C., and Gallate, J. (2012). The function of the anterior temporal lobe: a review of the empirical evidence. *Brain Res.* 1449, 94–116. doi: 10.1016/j.brainres.2012.02.017

Yang, H., Susilo, T., and Duchaine, B. (2016). *The anterior temporal face area contains invariant representations of face identity that can persist despite the loss of right FFA and OFA*. Cerebral cortex: New York, NY, *26*, 1096–1107.

Zhao, S., Xiang, Y., Xie, J., Ye, Y., Li, T., and Mo, L. (2017). The positivity bias phenomenon in face perception given different information on ability. *Front. Psychol.* 8:570. doi: 10.3389/fpsyg.2017.00570

Zillman, D., and Cantor, J. R. (1977). Affective responses to the emotions of a protagonist. *J. Exp. Soc. Psychol.* 13, 155–165. doi: 10.1016/S0022-1031(77)80008-5

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 01 December 2022
DOI 10.3389/fpsyg.2022.980261

Check for updates

OPEN ACCESS

EDITED BY
Luca Tummolini,
National Research Council, Italy

REVIEWED BY
Jan Ketil Arnulf,
BI Norwegian Business School, Norway
Jean-Marie C. Chevalier,
Université Paris-Est
Créteil Val de Marne, France

*CORRESPONDENCE
Frank Zenker
fzenker@gmail.com

SPECIALTY SECTION
This article was submitted to
Theoretical and Philosophical Psychology,
a section of the journal
Frontiers in Psychology

RECEIVED 28 June 2022
ACCEPTED 19 October 2022
PUBLISHED 01 December 2022

CITATION
Witte EH, Stanciu A and Zenker F (2022)
Predicted as observed? How to identify
empirically adequate theoretical
constructs.
Front. Psychol. 13:980261.
doi: 10.3389/fpsyg.2022.980261

COPYRIGHT
© 2022 Witte, Stanciu and Zenker. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Predicted as observed? How to identify empirically adequate theoretical constructs

Erich H. Witte[1], Adrian Stanciu[2] and Frank Zenker[3]*

[1]Institute for Psychology, University of Hamburg, Hamburg, Germany, [2]Data and Research on Society, GESIS-Leibniz Institute for the Social Sciences, Mannheim, Germany, [3]Department of Philosophy, Boğaziçi University, Istanbul, Turkey

The identification of an empirically adequate theoretical construct requires determining whether a theoretically predicted effect is sufficiently similar to an observed effect. To this end, we propose a simple similarity measure, describe its application in different research designs, and use computer simulations to estimate the necessary sample size for a given observed effect. As our main example, we apply this measure to recent meta-analytical research on precognition. Results suggest that the evidential basis is too weak for a predicted precognition effect of $d = 0.20$ to be considered empirically adequate. As additional examples, we apply this measure to object-level experimental data from dissonance theory and a recent crowdsourcing hypothesis test, as well as to meta-analytical data on the correlation of personality traits and life outcomes.

KEYWORDS

crowdsourcing hypothesis test, dissonance theory, empirical adequacy, Paul Meehl, meta-analysis, personality research, precognition, theory construction

> "*I am deliberately setting aside statistical significance testing, or the setting up of confidence intervals* […]"
>
> (Meehl, 1990; p. 128)

Introduction

As classical empirical findings fail to replicate and empirical studies often prove to be poorly conducted (Gervais, 2021; Nosek et al., 2022), the *replication crisis* or *confidence crisis* presents a major impasse for behavioral science (Fleck, 1935; Kuhn, 1962). While the motives for employing questionable research practices (Gelman and Carlin, 2014; Gelman, 2018) and the limitations of research methods (Kerr, 1998) are increasingly better understood, most reform proposals today recommend *transparency measures* (e.g., study pre-registration or registered replications; Fiedler and Prager, 2018; Klein et al., 2018). Less frequently addressed is that scientific progress requires good theoretical constructs (Meehl, 1978; Gigerenzer, 1998; Miłkowski et al., 2019; Muthukrishna and Henrich, 2019; Oberauer and Lewandowsky, 2019; Eronen and Romeijn, 2020; van Rooij and Baggio, 2020; Cornelissen et al., 2021; Eronen and Bringmann, 2021; Gervais, 2021; Irvine, 2021).

A good theoretical construct minimally allows for an empirically adequate prediction. A theoretical construct "is empirically adequate exactly if what it says about the observable things and events in the world is true—exactly if it "saves [or captures] the phenomena" (van Fraassen, 1980, p. 12). In the context of experimental research, this means that the effect that is predicted by a theoretical construct must be sufficiently *similar* to a relevant observed effect.

The development of an empirically adequate construct depends on high-quality observations. But even observations of the highest quality cannot automatically generate a theoretical construct that offers a *non-circular* justification for why a future event occurs as predicted.[1] Because a theoretical construct must deductively entail its prediction *before* observations are made, a non-circular approach to predicting a phenomenon of interest thus requires a *deductive* approach to the development of empirical adequate theoretical constructs (Popper, 1959; Lakens, 2013; Lakens et al., 2018).

We begin by summarizing why the empirical adequacy of a theoretical construct should be evaluated *independently* of statistical elements (Meehl, 1990; p. 128) and review the shortcomings of extant evaluative approaches. To this end, we propose a new formal measure that is independent of statistical elements, thus enabling a *direct* comparison between theory and observation. The intended application for this measure is theory construction. To demonstrate its use value, we exemplarily evaluate recent meta-analytical findings on *precognition* (Bem et al., 2016). Additional examples, as well as a description of how this measure can be applied under various research designs, are provided in Supplementary Appendix S1, S2.

Summary

Evaluating whether a theoretical prediction agrees with observations requires a theory-accommodating approach. But if this approach combinates theoretical and statistical aspects, then the evaluative outcome depends on the *variance* of error-prone observations. Consequently, one cannot be sufficiently certain about the accuracy of observations to which the theoretical prediction is compared. Since this uncertainty transfers to the evaluative outcome, the question of whether a theoretical prediction agrees with observations should be addressed *independently* of how observations vary (Meehl, 1990, 1992, 1997).

Yet the opposite holds if a standardized effect size measure such as Cohen's $d = (m_1 - m_0)/s$ is used to quantify the observations. This measure combines the observed *mean difference* $(m_1 - m_0)$ with the statistical element of the observed *standard deviation* (s). A theoretical construct, however, predicts only $(m_1- m_0)$, yet not s. This makes a standardized effect size measure an inappropriate formal tool to evaluate the empirical adequacy of a theoretical construct.

A theoretical construct contrasts most starkly with an inductive generalization that states a *directional* hypothesis. Because a directional hypothesis is informative only relative to its inductive basis, it can merely "predict" the pattern of past observations it subsumes. A theoretical construct, by contrast, is informative beyond this basis (see our note 1). Moreover, the construct must predict future observations not as a directional but as a *point-specific* effect. Otherwise, one simply cannot evaluate whether the theoretically predicted mean agrees with the observed mean.

Shortcomings of the inductive strategy

Standard deviation

The observed standard deviation (s) is a measure of the variance of observations. The observed variance depends on the

[1] The creative process of developing a theoretical construct is what C.S. Peirce called an *abduction* over past observations (Peirce, 1931–1958). Jointly with initial and boundary conditions, a theoretical construct allows for a *deduction* of a theoretical prediction about possible future observations. And a (dis-)confirmation of this prediction by *new* observations that are (in-)consistent with it relies on a testing process that the late C.S. Peirce called *induction*. Abduction thus is "the process of forming an explanatory hypothesis [and is] the only logical operation which introduces any new idea," whereas "deduction merely evolves the necessary consequences of a pure hypothesis," while induction "does nothing but determine a [truth] value" (Peirce, 5.171).
Importantly, the information content of a theoretical construct exceeds that of an *inductive generalization* (e.g., a mathematical function stating an observed law-like regularity) that *descriptively* subsumes past observations. Theoretical constructs acquire this excess content by featuring at least one *theoretical entity* that is not presupposed by the observational theory employed to make past observations (see Andreas, 2021, and our Supplementary Appendix S2, dissonance theory). Behavioral scientist, however, who likewise develop theoretical constructs based on past observations, regularly fail to acknowledge, and to perform, what Hempel (1988) called *theoretical (or inductive) ascent*, i.e., "[…] a transition from a data sentence expressed in [an antecedent vocabulary] V_A to a theoretical hypothesis […]" (p. 150) that is "formulated with the help of a theoretical vocabulary, V_C, whose terms refer to the kinds and characteristics of the *theoretical* entities and processes in question" (p. 147, *italics added*) which themselves are the products of *abduction*.
Without theoretical ascent, therefore, the information content of a *non*-genuine theoretical construct is at most as large as that of an inductive generalization. The main consequence is that the act of predicting *future* observations based on the inductive generalization that a *non*-genuine theoretical construct is, runs straight into Hume's problem of induction (Hume, 1739): a non-pragmatic justification for a prediction of future observations based on an inductive generalization of past observations (*sans* theoretical ascent) presupposes that induction is a *valid* mode of reasoning. But this inference is circular (Henderson, 2020).

extent to which an empirical setting is subject to uncontrolled (random) influences. Other things being equal, empirical settings that are more rigorously controlled for (random) influences go along with reduced observed variance, i.e., a smaller *s*. Compared to a less rigorously controlled setting, therefore, the value of Cohen's *d*-measure increases.

Since the observed standard deviation quantifies the variance of error-prone observations, an observed effect must be related to a probability distribution. This process is known as *standardization*. Once standardized, the observed effect becomes a statistic of an entire sample of observations that can no longer be related *directly* to a theoretically predicted effect. With a standardized observed effect, therefore, one cannot evaluate the *similarity* between what a theoretical construct predicts and what a measurement instrument records. Instead, one evaluates the relative position of statistically transformed measurement scores on a measurement scale against a random distribution.

A statistical test relies on the observed standard deviation to evaluate whether the observed effect differs statistically significantly from a null hypothesis. A *t*-test, for instance, can often show that a *large* difference between the observed means in the experimental and the control group is statistically significant. But the standard deviation *combines* several causes that contribute to the observed variance (e.g., the sample selection process, the experimental implementation, the validity and the reliability of the in- and dependent variables, and the random influences on an empirical setting). Thus, a theoretically predicted and an observed effect may well agree. But if the observed effect depends on the observed standard deviation, then its statistical significance is an *insufficient* criterion to evaluate a theoretical construct as empirically adequate.

Parameter estimation

Parameter estimation is an *inductive* strategy to separate systematic patterns from non-systematic noise in data. A parameter operates at the level of statistics rather than the level of measurement. 'Parameter' thus refers not to the properties of observations but those of data (e.g., their central tendency as measured by the mean, or the strength of associations between variables as measured by correlation or regression coefficients). Since data provide the basis for a parameter estimate, its accuracy is informed by statistical procedures that evaluate the parameter against the observed variance. The latter results from the variation of behavioral responses and measurement shortcomings. A given measurement instrument, therefore, captures both a relevant phenomenon *and* random influences (e.g., due to participants' salient memories, chronic moods, or even the weather).

This leads to three complications in estimating a parameter accurately. First, since perfectly error-free observations are impossible, the accuracy of a parameter must be evaluated against the observed variance by using statistical procedures (that rely on a significance level α and an associated probability level *p*). Such procedures are often subjective and need not be reliable (see *p*-harking, Kerr, 1998; *p*-hacking, Simmons et al., 2013). Crucially, statistical procedures cannot distinguish whether the observed variance results from measurement shortcomings or rather from uncontrolled (random) influences on an empirical setting.

Second, what matters for scientific discovery is the size of the parameter estimate. For instance, a small observed mean difference between people's political orientation that varies with color preferences presumably *fails* to be a substantially meaningful finding. Whereas a similarly small observed difference that varies with cultural background presumably would be substantially meaningful. This finding, however, should be further explored only if it is sufficiently large. But recent *meta*-meta-analyses (Olsson-Collentine et al., 2020; Schauer and Hedges, 2020; Linden and Hönekopp, 2021) strongly suggest that individual published studies across different behavioral science domains typically report observed object-level effects that are small and homogenous (read: small *d*, small *s*) or medium-to-large and heterogeneous (read: large(r) *d*, large *s*). A small observed variance thus tends to go along with a small observed mean effect. Whereas the findings of individual object-level studies that are sufficiently large to be further explored go along with a large observed variance. This necessarily results in a *vague* impression of the parameter that an empirical adequate theoretical construct would have to predict.

Third, a parameter estimate is useful for theory construction only if its inductive basis accurately captures an observed effect in a relevant population. Considerations of test-power and sample representativity dictate the use of sufficiently large samples to discover systematic behavioral patterns (law of large numbers). In small samples, by contrast, these patterns are likely truncated by uncontrolled (random) influences, resulting in inaccurate parameter estimates. Generally, large samples allow for more accurate parameter estimates if the underlying distribution of observations is uniform.

Among the widely used tools to estimate parameters are Cohen's *d*-measure, confidence intervals, and tools that rely on inductive model fitting and probabilistic distributions.

Cohen's *d*-measure

The goal of null-hypothesis significance testing is to determine whether an observed object-level effect differs significantly from a random effect. Relative to a predefined significance level α and an associated probability level *p*, the statistical significance of an observed effect indicates the probability of observing this effect under the null hypothesis. But this says nothing about whether the null or the alternative hypothesis is true or whether the observed object-level effect is relevant for theory construction. For theory construction, therefore, the statistical significance of an observed object-level effect is merely a necessary criterion. In addition, publications should also report the observed object-level effect's size.

Among the available tools to calculate the observed effect size, standardized effect size measures are often preferred because they weigh the observed effect by the observed variance, thus providing a robustness check for the observed effect. As one of the most widely used measures in behavioral science (Schäfer and Schwarz, 2019), for instance, Cohen's standardized d-measure $d = (m_1 - m_0) / s$ (Cohen, 1977) weighs the observed mean difference ($m_1 - m_0$) between the experimental (m_1) and the control group (m_0) by the pooled standard deviation in both groups (s). It should be easy to see that, if ($m_1 - m_0$) is constant, then the d-value is sensitive to the observed variance captured by s.

Even if an experimental study that relies on the d-measure would report a very *large* statistically significant effect, this is insufficient to motivate the development of a theoretical construct for it. To be theorized, after all, is the *true* parameter, rather than its ratio to the observed variance. The main challenge thus is to tease apart the causes that contribute to the observed variance (see above). Standardized effect size measures, however, simply cannot meet this challenge, making them inappropriate tools for theory construction research. Therefore, an additional layer of scrutiny must address the confidence that an inductive parameter estimates the true parameter.

Confidence intervals

A true parameter can be estimated with perfect accuracy only in theory. In praxis, (random) influences or measurement instrument shortcomings render a perfectly accurate parameter estimate unlikely. One can nevertheless state the parameter's *expected* accuracy using a *confidence interval* (CI), the width of which depends on the level of significance α. To determine the CI, one simultaneously considers the observed mean difference, the observed variance, the level of significance, and the sample size. This is formally given as $CI = d \pm z \times \left(s / \sqrt{n} \right)$.

Like Cohen's d-measure, however, a CI cannot determine whether a true effect (e.g., the mean difference between two groups in a population) was estimated accurately because also a CI combines the mean difference with the statistical element s. Thus, the observed variance once again results in a *vague* impression of the parameter. Generally, unless the causes that contribute to the observed variance can be teased apart, vague observations will undermine theory construction research. And the one possible way of teasing these causes apart is to increase the sample size.

Inductive model fitting

Using inductive model fitting, researchers can address the complexity of human behavior by statistically modeling the associations between two or more estimated parameters, followed by testing the statistical model against a random model. Using various indexes (e.g., the Comparative Fit Index (CFI) or the Root Mean Square Error of Approximation (RMSEA)), a finite set of observations is compared against a class of statistical models (see the special issue on model selection, Myung et al., 2000; Burnham and Anderson, 2004). The model that best describes the data is said to be *identified* in the population (Bollen et al., 2010).

Inductive model fitting presupposes a reconstruction of the variance–covariance structure in the data. But fitting a statistical model to data inherits all attributes of the data (including errors due to measurement instrument shortcomings, uncontrolled random influences, non-uniform distributions, or outliers). So, although inductive model fitting improves over the estimation of a single parameter, its use-value for theory construction primarily depends on the quality of the data. Even the best-fitted model, however, cannot *unequivocally* tell meaningful data patterns from patterns owed to measurement instrument shortcomings or uncontrolled (random) influences. This holds regardless of whether the estimated parameter is statistically significant or whether the effect size is large. All an inductively fitted model can tell is whether data are described well.

Since model fitting is an *iterative* strategy, moreover, some parameters must be estimated before others, so that the associations between parameters can be specified to obtain a data-fitting model. The identification of the parameters that are to be estimated first would ideally rely on theoretical considerations. But when researchers fit a model to data, they instead often rely on p-harking or p-hacking strategies.

Bayesian probabilistic distributions

In the Bayesian approach to parameter estimation, the known probability of past observations is assumed to estimate the probability of (predicted) future observations. A theoretical construct can thus be evaluated based on the prior probability of a statistical model (Wagenmakers and Farell, 2004). The observed variance is here captured by the assumption that the theoretical construct is itself subject to variation. So, rather than evaluating the agreement between data and a single statistical model, Bayesians evaluate the agreement between data and a *distribution* of possible statistical models.

A theoretical construct is thus specified not as a single parameter, but as one that is embedded in a prior probability distribution (e.g., a normal or a Cauchy distribution). Of course, if this prior probability distribution accurately captures the true parameter, then a theoretical construct that is specified as a probability distribution may be useful for theory construction. What the *true* probability distribution is, however, one can never know. A Bayesian parameter estimate, therefore, depends not so much on the quality of the data, but more on a researcher's (subjective) assumptions about the prior probability distribution (see Krefeld-Schwalb et al., 2018).

Since the theoretical construct is more likely to be associated with an upper and a lower probability bound than with a unique probability, the Bayesian approach to parameter estimation corresponds—except for the distribution of possible theoretical parameters—to the specification of a theoretical construct as an *interval* hypothesis (i.e., a two-point-hypothesis). Because the endpoints of this interval represent two distinct theoretical parameters, each endpoint must be *separately* evaluated against data. But the possibility of a separate evaluation of two theoretical parameters also shows that there is no genuine need to distribute them. After all, if the (subjective) *a priori* probabilities of both theoretical parameters are independent, then as one parameter is assigned probability 1, the other can be assigned probability 0.

Toward a deductive strategy: Paul Meehl's corroboration index

In the context of theory construction research, probably the first in behavioral science to recognize a problem in relating the theoretically predicted effect to the sample statistic *s* was Meehl (1990). Against the background of Lakatos' (1978) "core vs. protective belt"-model of empirical theories—which recognizes that making suitable adjustments to the protective belt can (in principle forever) deflect the empirically inadequate predictions that constitute a theory's *falsification instances* away from the core—Meehl argued that a formal measure for the empirical adequacy of a theoretical construct should *ignore s*.

> "To construct a crude [corroboration-]index of a theory's [predictive] track record, one first amends the earlier Popper to the later Popper by shifting emphasis from falsification to verisimilitude. […] Meanwhile, we require of a candidate index that it somehow reflect how bad a numerical "miss" the experimenter chalks up against [the theory] T. […] We are examining the relationship between T and its track record in predicting numerical values of [a hypothesis] H, *ignoring the stochastic slippage* between H and the data set that is the main concern of the statistician."

(Meehl, 1990; p. 128)

Meehl's corroboration index (C_i) is the following:

$$C_i = (Cl) \times (In) \quad (1)$$

where Cl = the closeness of observed data to the theoretical prediction;

In = the intolerance of the theory (e.g., the standardized precision of a prediction).

These terms can be expanded:

$$Cl = 1 - (D / S) \quad (2)$$

where D = the deviation of observed data from the tolerance interval of the theory;

S = "Spielraum," i.e., the expected range of observed data regardless of whether the theory is true; and

$$In = 1 - (I / S) \quad (3)$$

where I = the interval tolerated by the theory (or the raw precision of a theoretical prediction).

For a given experiment, the index C_i is the product of the *closeness* of the data to the theoretical prediction (Cl) and the *intolerance* of a theory (In). Thus, large values of C_i are expected for an empirically adequate theoretical construct and small values of C_i for an empirically inadequate one. Although several critics considered the C_i measure overly complex (see the special issue of *Psychological Inquiry*, including Meehl 1990), Meehl (1992) rightly replied that formal measures are needed to develop empirically adequate theoretical constructs. Yet, Meehl's key insight—that a formal measure to evaluate the empirical adequacy of a theoretical construct should *ignore* the statistical element *s*—further awaits uptake. Researchers instead continue to rely on statistical considerations (e.g., CIs, *t*, *d*, etc.) or on model-fitting approaches that combine theoretical with statistical elements.

Heeding Meehl's insight, we propose the *similarity index* I_{SIM} as an alternative formal measure, one far simpler than C_i.

The similarity index

As we saw, if a parameter is induced from an interval of observations, then the parameter captures the uncontrolled (random) influences on an empirical setting that are represented by *s*. Although this parameter may (misleadingly) be referred to as a theoretical construct, this construct is as *vague* as the underlying interval of observations is wide. An inductive parameter, therefore, is at most as informative as a two-point, directional alternative hypothesis (H_1). But a directional alternative hypothesis cannot stand in the *one-to-one* relation between prediction and observation that is required to evaluate whether a theoretical construct is empirically adequate (Klein, 2014; Szucs and Ioannidis, 2017; Gelman, 2018). Only a point-specific theoretical construct can do so.

For this reason, Meehl (1990) argued that the evaluation of the empirical adequacy of a theoretical construct should ignore *s*. Once the evaluation is independent of *s*, it pertains only to the *similarity* between a predicted and an observed mean difference in a sample. This is precisely what the similarity index I_{SIM} captures (see 4).

$$I_{\text{SIM}} = \frac{|m_{\text{THEO}} - m_0|}{|m_1 - m_0|} = \text{ES}_{\text{THEO}} / \text{ES}_{\text{OBS}} \quad (4)$$

ES, effect size.

m_{THEO}, the theoretically predicted mean.

m_1, the observed mean in the treatment group.

m_0, the observed mean in the control group.

$m_{\text{THEO}} - m_0$, the theoretically predicted mean difference (ES_{THEO}).

$m_1 - m_0$, the empirically observed mean difference (ES_{OBS}).

A formal measure for the empirical adequacy of a theoretical construct should satisfy several criteria that are relevant to theory construction. First, an experimentally observed phenomenon must be independent of the measurement scale that a given measurement instrument presupposes. Second, any two phenomena that are recorded on distinct measurement scales must remain comparable. Third, observations must remain stable under theoretically plausible transformations.

But if different measurement scales are made comparable by a transformation into z-values, then recourse to the inductive element s entails that the measurement quality of the empirical setting is retained. A z-transformation thus *inherits* information originating from the uncontrolled (random) influences on an empirical setting. This is problematic for theory construction research because, given that s as a property of observations *lacks* a theoretical counterpart, recourse to s "blurs" the evaluation of the empirical adequacy of a theoretical construct.

I_{SIM} uses a transformation that avoids s. The comparability of observations that are recorded on different measurement scales is guaranteed because a ratio of differences is invariant under the addition of a constant or multiplication by some factor.[2] I_{SIM} also guarantees that the direction of the observed effect can be interpreted. This matters for evaluating whether the observed effect leans toward the experimental or the control group. If the direction of the observed effect and the theoretically predicted effect agree, then I_{SIM} is invariant concerning the order of means. That the same mathematical signs (+, –) now appear in the numerator and the denominator of I_{SIM} can be neglected. Whereas if the direction of the observed effect and that of the theoretically predicted effect differ, then distinct mathematical signs indicate that the prediction fails to agree with observations. In this case, I_{SIM} is set to 0.

Using I_{SIM}, the theoretically predicted effect can thus be compared *directly* to the observed effect. A direct comparison should arguably also apply if a theoretically predicted effect is compared to a meta-analytically estimated population effect that is aggregated from the results of independent replication studies. But the opposite is the case if this comparison relies on a standardized effect size measure such as Cohen's d, which is widely used for this purpose today. Sometimes, indeed, the observed d-value simply stands in for the estimated population effect.

The intended application for I_{SIM} is a rigorously controlled empirical setting where participants are randomly allocated to the experimental and the control group, respectively are randomly selected as study participants in a correlational study.[3] Since the use of this kind of setting to evaluate the empirical adequacy of a *directional* H_1 undermines all efforts at controlling the setting, a rigorously controlled empirical setting should exclusively serve to evaluate the high-risk prediction that only a point-specific theoretical construct can offer.

The similarity between theory and observations

The agreement between a theoretical prediction and observations is *perfect* if the ratio between both is one, i.e., $\mathrm{ES}_{\mathrm{THEO}} / \mathrm{ES}_{\mathrm{OBS}} = 1.00$. A perfectly empirically adequate prediction, however, is a strong idealization because even the most rigorously controlled empirical setting is subject to some uncontrolled (random) influences and errors. So, even if a theoretical construct predicts a population effect perfectly (i.e., $\mathrm{ES}_{\mathrm{THEO}} = \mathrm{ES}_{\mathrm{POP}}$), a measurement instrument with imperfect reliability or random influences on an empirical setting do entail that the observed effect will be "blurred." A formal measure for the empirical adequacy of a theoretical construct, therefore, can only *approximate* the agreement between a theoretical prediction and observations.

Analytically, the agreement between a theoretical prediction and observations varies between a *match* ($I_{\mathrm{SIM}} = 1.00$) and a *mismatch* in one of two directions ($I_{\mathrm{SIM}} = 0$ and $I_{\mathrm{SIM}} >> 1$). The reason for a mismatch—namely whether the theoretical construct predicts an empirically inadequate effect or whether the observed effect is subject to random influences—can be teased out by collecting additional data, i.e., by increasing the sample size n. If the values of I_{SIM} cluster around 1 as n increases, this indicates that the theoretically predicted effect approximately matches a relevant population effect (law of large numbers). As the observed effect thus progressively converges onto the population effect ($\mathrm{ES}_{\mathrm{OBS}} = \mathrm{ES}_{\mathrm{POP}}$), it can eventually be excluded that random influences account for the observations. Thus, one gains evidence that the theoretically predicted effect is *empirically adequate*. This case is perfect for theory construction because the theoretical construct can be adopted into a theory.

Whereas if values of I_{SIM} never cluster around 1 as n increases, then the theoretical prediction is *empirically inadequate*. This means one gains evidence that the theoretically predicted effect misrepresents the population effect, wherefore the theoretical construct requires adjustment. Subsequently, a new theoretically predicted effect must be separately evaluated using new observations.

The similarity interval

Defining the range of acceptable deviations from a perfect match requires an interval of the form $[x < I_{\mathrm{SIM}} = 1.00 < y]$. The purpose of this *similarity interval* (SI) is distinct from that of a *confidence interval* (CI). When a population effect ($\mathrm{ES}_{\mathrm{POP}}$) is estimated from observations, a CI handles randomly distributed "noise" in an empirical setting by stating the interval within which $\mathrm{ES}_{\mathrm{POP}}$ is expected to lie to some predefined probability (see the section *Parameter Estimation*). The SI, by contrast, differentiates between evidence for and against the empirical adequacy of a theoretical construct by stating the probability that the theoretically predicted effect is similar to observations if a study is *repeated* numerous times.

[2] With 'a' for the *origin* of the scale (normalization), 'u' for the *unit* of the scale (standardization), and 'x, y, z' for arbitrary *measurement values*, the ratio $[(x+a)\,u - (y+a)\,u] \,/\, [(z+a)\,u - (y+a)\,u] = (x - y) \,/\, (z - y)$ is invariant for all values of a and u.

[3] Such rigor often cannot be achieved. Researchers in personality psychology, for instance, typically cannot randomly allocate study participants according to their personality characteristics. The object of inquiry, therefore, are not treatment effects but correlations between variables. These correlations can nevertheless be generalized to a population if a sample is representative of it (Kish, 1965).

The SI is motivated by two constraints. First, an empirically adequate theoretical construct must neither grossly under- nor grossly over-predict the population effect ($ES_{THEO} \cong ES_{POP}$). Second, provided the first constraint holds, if the theoretically predicted effect keeps approximating the observed effect as the number of study repetitions increases, then the theoretically predicted effect becomes increasingly more promising as a parameter for theory construction because the prediction remains empirically adequate.

The SI particularly facilitates the identification of a *preliminary* match between a theoretically predicted effect (ES_{THEO}) and an observed effect (ES_{OBS}), because an I_{SIM}-based evaluation is fallible—future studies may lead to an opposite evaluation. We define a preliminary match using an SI with bounds of [0.80;1.20]. If the I_{SIM} value lies within these bounds, then the theoretical prediction is preliminarily empirically adequate. The bounds [0.80;1.20] are informed by 10,000 simulated study repetitions (see the section *Simulated Data and Results*). For instance, given $n_0 = n_1 = 1{,}000$ participants, our simulations show that if the population effect is a *medium* effect, $ES_{POP} = 0.50$, then I_{SIM}-values fall within this SI in approximately 99% of 10,000 study repetitions. And, given $n_0 = n_1 = 100$ participants in each study condition, if the population effect is a *large* effect, $ES_{POP} = 1.00$, then I_{SIM}-values fall within the SI in approximately 95% of 10,000 repetitions.

Since a *small* sample suffices to detect a *large* population effect under small error-rates, whereas detecting a *small* population effect requires a *large* sample, the application of a 99%-SI to the small to medium effects that are normally observed in behavioral science would require unrealistically large samples (Linden and Hönekopp, 2021). Given the conventional error rate of 5%, however, already a 95%-SI can suffice as an evidence-based criterion to decide whether a theoretical construct can be accepted as empirically adequate, whether it should be improved, or whether additional data should be collected.

Simulated data and results

If simulations approximate the universe of possible observed effects, they are useful to explore the stability of effects that real studies would observe (see Morris et al., 2019). Real observations are made in samples drawn from some population of interest. But researchers typically cannot access the entire population, neither in real life nor in simulations. To account for the ultimately unknown observed variance, real observations are treated statistically as a t-distribution, which is sensitive to n. As n increases, a t-distribution approximates the normal distribution that is expected for a population (central limit theorem).

We therefore simulated data from t-distributions in a universe of study settings that comprises 10,000 repeated individual studies of the same effect. A study setting is characterized by the means observed in the control (m_0) and the experimental group (m_1) and by the sample size ($n_0 = n_1$). All simulations were conducted in R (R Core Team, 2021) using the packages tidyverse (Wickham et al., 2019), dplyr (Wickham et al., 2021), and effsize (Torchiano, 2020).

In the first of two basic scenarios, where the theoretically predicted effect *matches* the population effect ($ES_{THEO} = ES_{POP}$), the sample size of a study setting was $n_0 = n_1 = 20$, 30, 50, 100, 300, or 1,000. In the control group the observed mean was null ($m_0 = 0$) and in the experimental group $m_1 = 0.20$, 0.50, 0.80, 1.00, 1.20, 1.40, 1.60, 1.80, or 2.00. In this way, we simulated 54 study settings times 10,000 repetitions, calculating the similarity index I_{SIM} separately for each repetition of a study setting (see formula 4). For the percentages of I_{SIM}-values falling inside and outside the similarity interval SI, see Table 1 and Figure 1.

Findings are consistent with the claim that an empirical adequate theoretical construct is associated with values of I_{SIM} that fall inside the SI [0.80;1.20]. For example, values of I_{SIM} fall inside this SI in approximately 95% of study repetitions if the sample size is $n_0 = n_1 = 100$ and if $(m_1 - m_0) = 1.00$. In contrast, values of I_{SIM} fall inside this SI in approximately 67% of study repetitions given the same sample size and a smaller effect of $(m_1 - m_0) = 0.50$. This suggests that $n_0 = n_1 = 100$ suffices to evaluate a *large* theoretically predicted effect as preliminarily empirically adequate, whereas evaluating a *small* or *medium* theoretically predicted effect requires a considerably larger sample.

The second scenario, where the theoretically predicted effect *failed* to match the population effect ($ES_{THEO} \neq ES_{POP}$), examined how false positive and false negative predictions fare in our simulated universe of study repetitions. A false positive prediction occurs if the theoretically predicted effect is mistakenly identified as matching the population effect. And a false negative prediction occurs if the value of I_{SIM} falls outside the SI despite the theoretically predicted effect matching the population effect. In this scenario, we simulated four study settings where the theoretically predicted effect varied from small to large, and the population effect was either over- or underestimated. Notice that the relevant quantity to guide the identification of an empirically adequate theoretical construct here is not the *absolute* probability of detecting an empirically (in-)adequate prediction, but the *difference* between the probabilities of detecting one or the other kind of prediction.

In each of the four study settings, the sample size was $n_0 = n_1 = 20$, 30, 50, 100, 300 or 1,000. In two of the four study settings, the theoretically predicted effect *overestimates* the population effect. Setting 1 simulated data from t-distributions representing a population effect of $ES_{POP} = 0.20$, whereas the theoretically predicted effect was $ES_{THEO} = 0.50$. Setting 2 simulated data from t-distributions representing a population effect of $ES_{POP} = 0.80$, whereas the theoretically predicted effect was $ES_{THEO} = 1.00$. In the remaining two study settings, the theoretically predicted effect *underestimates* the population effect. Setting 3 simulated data from t-distributions representing a population effect of $ES_{POP} = 0.80$, whereas the theoretically predicted effect was $ES_{THEO} = 0.20$. Setting 4 simulated data from t-distributions representing a population effect of $ES_{POP} = 1.20$, whereas the

TABLE 1 True predictions [ES_{THEO}=ES_{POP} equals (m_{THEO} − m_0)=(m_{POP} − m_0)]: expected I_{SIM}-values for varying values of m_{THEO} and *n*.

$m_{THEO} = m_{POP}$	*n*	<0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1.0]	(1.0,1.1]	(1.1,1.2]	(1.2,1.3]	(1.3,1.4]	>(1.4
0.20											
	20	52.2	5.55	5.67	5.34	4.52	3.44	2.69	2.3	1.96	16.33
	30	45.89	6.64	6.43	5.7	4.84	4.14	3.61	2.78	1.92	18.05
	50	35.97	7.86	7.43	7.29	6.29	5.06	4.24	3.49	2.75	19.62
	100	20.84	8.4	9.97	9.98	9	7	5.3	4.39	3.46	21.66
	300	3.26	5.69	11.86	15.24	15.25	12.31	9.1	6.6	4.6	16.09
	1,000	0.04	0.71	5.62	17.65	25.46	21.84	13.84	6.94	3.87	4.03
0.50											
	20	18.04	8.61	10.18	10.67	8.86	7.93	6.07	4.77	3.71	21.16
	30	10.25	8.03	10.77	12.29	11.78	9.09	7.05	5.74	4.36	20.64
	50	3.32	6.03	10.57	14.8	15.65	12.71	9.16	6.33	4.28	17.15
	100	0.33	2.58	8.69	18.15	20.52	16.48	11.87	7.1	5.31	8.97
	300	0	0.04	1.96	15.78	32.69	28.31	13.69	5.02	1.51	1
	1,000	0	0	0	3.68	46.41	42.07	7.32	0.5	0.02	0
0.80											
	20	3.82	5.49	11.44	15.23	14.87	12.24	8.93	7.06	4.26	16.66
	30	1.1	4.02	10.5	16.51	18.03	15.48	10.74	7.25	4.96	11.41
	50	0.08	1.26	7.74	17.78	23.74	19.29	12.41	7.53	3.96	6.21
	100	0	0.05	2.42	16.08	30.82	26.69	14.27	6.06	2.24	1.37
	300	0	0	0.11	6.03	44.34	39.03	9.41	1.02	0.05	0.01
	1,000	0	0	0	0.28	49.5	49.18	1.04	0	0	0
1.00											
	20	1.19	3.96	10.87	16.73	17.35	15.2	10.68	6.71	4.87	12.44
	30	0.19	1.65	8.32	18.17	22.51	17.97	12.35	7.55	4.49	6.8
	50	0.01	0.29	4.33	17.45	28.02	22.9	13.84	7.24	3.09	2.83
	100	0	0	0.8	11.7	36.43	32.75	13.23	3.98	0.82	0.29
	300	0	0	0	2.75	48.25	43.03	5.77	0.2	0	0
	1,000	0	0	0	0.04	49.99	49.84	0.13	0	0	0
1.20											
	20	0.17	2.07	8.31	18.05	21.51	17.93	12.61	7.18	4.25	7.92
	30	0.03	0.57	5.43	18.69	25.74	21.59	13.15	7.29	3.51	4
	50	0	0.02	2.03	15.88	31.69	27.75	14.34	5.33	1.87	1.09
	100	0	0	0.25	9.25	39.33	37.34	11.53	1.95	0.33	0.02
	300	0	0	0	1.15	48.05	47.62	3.12	0.06	0	0
	1,000	0	0	0	0	49.58	50.4	0.02	0	0	0
1.40											
	20	0.02	1.03	6.74	17.44	24.80	20.73	13.02	7.28	4.04	4.90
	30	0	0.21	3.40	17.22	28.40	25.32	13.83	6.60	2.90	2.12
	50	0	0.01	0.95	13.18	36.19	30.6	13.48	4.04	1.15	0.4
	100	0	0	0.01	6.21	43.14	39.94	9.67	0.99	0.03	0.01
	300	0	0	0	0.46	49.71	48.36	1.47	0	0	0
	1,000	0	0	0	0	49.94	50.06	0	0	0	0
1.60											
	20	0	0.39	4.62	17.25	27.52	23.28	13.94	6.78	3.32	2.90
	30	0	0.08	2.02	15.50	33.03	27.58	13.67	5.17	1.96	0.99
	50	0	0	0.44	10.87	38.88	34.33	12.22	2.79	0.38	0.09
	100	0	0	0	4.06	45.70	42.55	7.25	0.43	0.01	0
	300	0	0	0	0.1	50.43	48.8	0.67	0	0	0
	1,000	0	0	0	0	50.4	49.6	0	0	0	0
1.80											
	20	0	0.19	2.83	16.56	30.69	25.19	14.10	6.30	2.43	1.71

(Continued)

TABLE 1 (Continued)

$m_{THEO} = m_{POP}$	n	<0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1.0]	(1.0,1.1]	(1.1,1.2]	(1.2,1.3]	(1.3,1.4]	>(1.4
	30	0	0.01	0.94	13.79	35.09	30.50	13.77	4.43	1.06	0.41
	50	0	0	0.1	8.07	42.21	36.61	10.85	1.83	0.28	0.05
	100	0	0	0	2.26	48.66	43.88	4.98	0.22	0	0
	300	0	0	0	0.03	49.82	49.99	0.16	0	0	0
	1,000	0	0	0	0	49.72	50.28	0	0	0	0
2.00											
	20	0	0.11	2.01	15.78	32.42	27.64	14.13	5.23	1.82	0.86
	30	0	0	0.49	11.42	38.40	32.96	12.47	3.33	0.71	0.22
	50	0	0	0.07	6.51	44.16	39.04	9.13	1	0.08	0.01
	100	0	0	0	1.45	48.14	46.84	3.49	0.08	0	0
	300	0	0	0	0	50.68	49.21	0.11	0	0	0
	1,000	0	0	0	0	50.62	49.38	0	0	0	0

Cells state the percentages of I_{SIM}-values falling within specific I_{SIM} intervals for various sample sizes (n), based on 10,000 simulations per row; (, value not included;], value included.

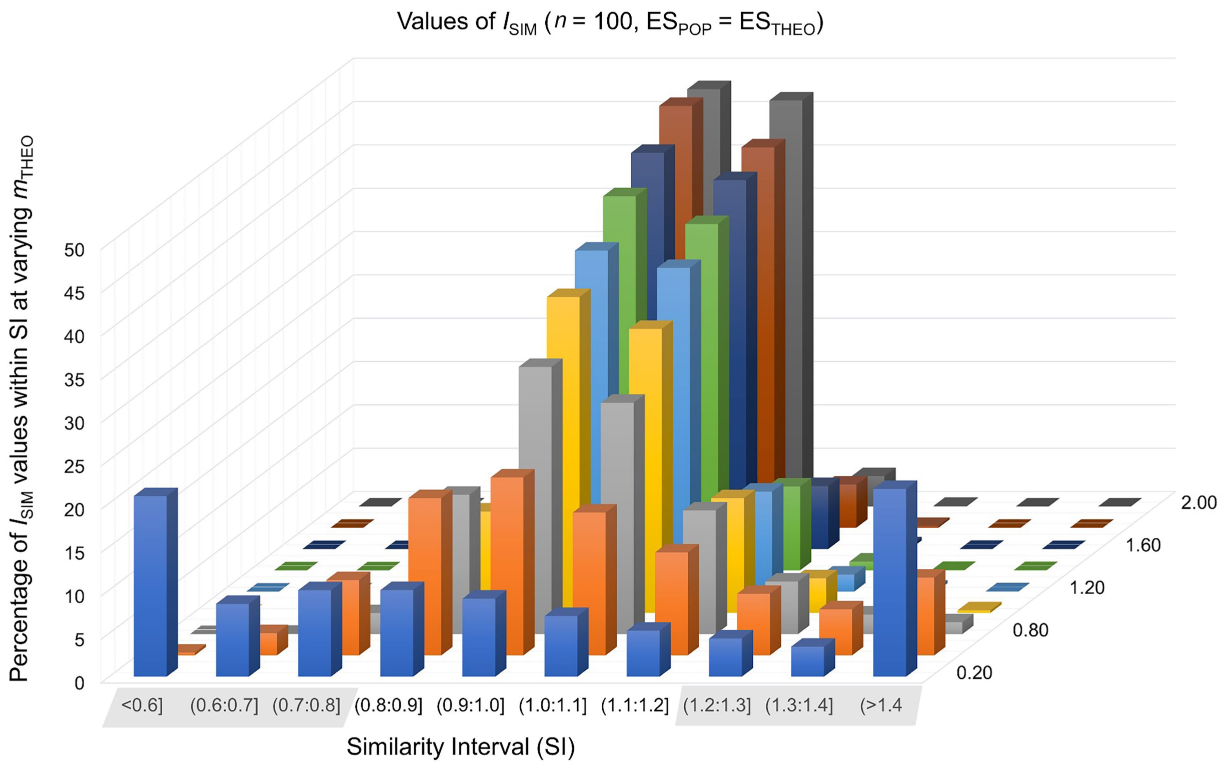


FIGURE 1
Values of I_{SIM} were calculated in 10,000 simulated study-settings with $n_0 = n_1$ =100 under the assumption that the theoretically predicted effect matches the population effect. Each row of this graph represents different values of m_{THEO}.

theoretical effect was $ES_{THEO} = 1.00$. All four study settings were repeated 10,000 times. For the percentages of I_{SIM}-values falling inside and outside the SI, see Table 2 and Figure 2.

We first turn to cases where the theoretically predicted effect *overestimates* the population effect. Given a sample size of $n_0 = n_1 = 100$, values of I_{SIM} fall inside the SI in approximately 2% of repetitions of setting 1 ($ES_{POP} = 0.20$, $ES_{THEO} = 0.50$), compared to approximately 31% of repetitions of a study setting where the theoretically predicted effect matches the population effect ($ES_{THEO} = ES_{POP} = 0.20$). The 29% difference between false positive and true positives predictions increases as n increases (see Tables 1, 2). For the 2% of false positive predictions, the decision is clear: the theoretical construct requires adjustment. Whereas in case of the 31% true positive predictions, the identification of an empirically adequate construct would benefit from increasing n.

TABLE 2 False predictions ($ES_{THEO} \neq ES_{POP}$): expected I_{SIM}-values given discrepancies between ES_{THEO} ($m_{THEO} - m_0$) and ES_{POP} ($m_{POP} - m_0$) for varying n.

$m_{THEO} \neq m_{POP}$	n	<0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1.0]	(1.0,1.1]	(1.1,1.2]	(1.2,1.3]	(1.3,1.4]	>(1.4
0.50 ≠ 0.20											
	20	28.08	0.84	1.47	2.70	3.38	3.86	4.18	3.98	3.56	47.95
	30	22.41	0.40	0.94	1.37	2.53	3.09	3.95	4.19	3.88	57.24
	50	16.11	0.03	0.11	0.54	1.19	2.21	2.99	3.86	4.15	68.81
	100	7.47	0	0	0.01	0.13	0.45	1.04	1.93	3.29	85.68
	300	0.59	0	0	0	0	0	0	0.09	0.39	98.93
	1,000	0	0	0	0	0	0	0	0	0	100
0.50 ≠ 0.80											
	20	45.21	18.25	13.53	8.23	5.09	3.08	1.66	1.15	0.76	3.04
	30	44.90	21.76	15.61	8.48	4.16	2.19	0.78	0.67	0.33	1.12
	50	42.86	28.41	17.45	6.74	2.66	1.1	0.43	0.17	0.07	0.11
	100	38.55	39.89	17.09	3.7	0.6	0.13	0.01	0.02	0.01	0
	300	31.46	59.73	8.68	0.13	0	0	0	0	0	0
	1,000	19.03	80.32	0.65	0	0	0	0	0	0	0
1.00 ≠ 0.80											
	20	0.80	0.69	2.53	5.76	10.07	12.96	12.61	10.64	8.69	35.25
	30	0.14	0.16	0.97	4.04	9.52	13.54	14.83	13.22	11.21	32.37
	50	0.03	0	0.09	1.51	6.88	14.29	18.59	17.29	13.49	27.83
	100	0	0	0	0.11	1.99	11.39	22.93	24.65	17.87	21.06
	300	0	0	0	0	0.07	2.97	25.10	42.35	22.06	7.45
	1,000	0	0	0	0	0	0.03	14.63	68.47	16.37	0.50
1.00 ≠ 1.20											
	20	4.72	13.52	23.39	23.79	15.41	8.48	4.53	2.37	1.66	2.13
	30	1.99	11.73	27.04	27.82	17.23	8.00	3.36	1.59	0.64	0.60
	50	0.26	7.22	29.28	36.03	19.12	6.02	1.53	0.37	0.09	0.08
	100	0	2.13	28.44	50.06	17.3	1.87	0.19	0	0.01	0
	300	0	0.02	19.69	73.84	6.41	0.04	0	0	0	0
	1,000	0	0	5.93	93.82	0.25	0	0	0	0	0

Cells state the percentages of I_{SIM}-values falling within specific I_{SIM} intervals for various sample sizes *(n)*, based on 10,000 simulations per row; (, value not included;], value included; m_{POP}, the true value in the population (i.e., the numerator of I_{SIM}); m_{THEO}, the theoretical value in the population (i.e., the denominator of I_{SIM}).

Further, given a sample size of $n_0 = n_1 = 100$, values of I_{SIM} fall inside the SI in approximately 36% of repetitions of setting 2 ($ES_{POP} = 0.80$, $ES_{THEO} = 1.00$), compared to approximately 88% of repetitions of a study setting where the theoretically predicted effect matches the population effect ($ES_{THEO} = ES_{POP} = 0.80$). The 53% difference between false positive and true positive predictions increases as *n* increases. In both cases, however, the decision to adjust the theoretical construct requires considerably larger samples to clearly distinguish a true positive from a false positive prediction.

We now turn to cases where the theoretically predicted effect *underestimates* the population effect. Given a sample size of $n_0 = n_1 = 100$, values of I_{SIM} fall inside the SI in approximately 4% of repetitions of setting 1 ($ES_{POP} = 0.80$, $ES_{THEO} = 0.20$), compared to approximately 88% of repetitions of a study setting where the theoretically predicted effect matches the population effect ($ES_{THEO} = ES_{POP} = 0.80$). The 84% difference between true positives and false positive predictions arguably suffices to evaluate the theoretical construct as empirically inadequate.

Finally, given a sample size of $n_0 = n_1 = 100$, values of I_{SIM} fall inside the SI in approximately 69% of repetitions of setting 1 ($ES_{POP} = 1.20$, $ES_{THEO} = 1.00$), compared to approximately 97% of repetitions of a study setting where the theoretically predicted effect matches the population effect ($ES_{THEO} = ES_{POP} = 1.20$). The 28% difference between false positives and true positive predictions suggests that it is more likely that values of I_{SIM} fall inside the SI if the theoretical prediction matches the population effect than otherwise.

We proceed to exemplify the application of I_{SIM} with a case study. Additional examples are provided in Supplementary Appendix S2.

Case study: The psi-effect

The question of whether humans can cognize the future (aka *precognition* or *psi*-effect) has interested several scholars in psychology. The authors of the largest meta-analysis on the psi-effect to date (Bem et al., 2016), comprising 90 experimental

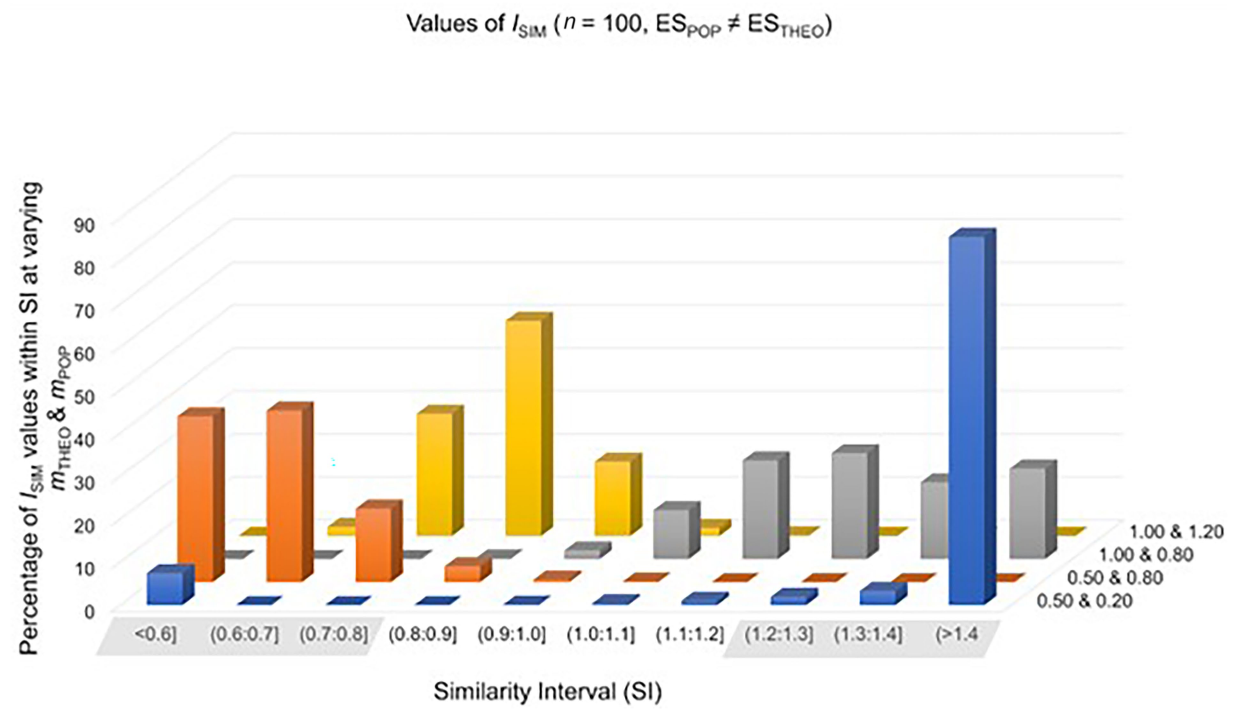


FIGURE 2
Values of I_{SIM} were calculated in 10,000 simulated study-settings with n_0=n_1=100 under the assumption that the theoretically predicted effect does not match the population effect. Each row of this graph represents different combinations of m_{THEO} and m_{POP}.

studies of which 51 are peer-reviewed (see Bem et al., 2016; Supplementary Table S1), claim to have obtained decisive evidence *for* a psi-effect. Whereas some concluded from this that the psi-effect is real (e.g., Cardena, 2018), others argued that Bem et al.'s (2016) meta-analytical data leave it too unlikely that the psi-effect is real (e.g., Witte and Zenker, 2017).

Across the 51 peer-reviewed object-level psi-studies, the observed effect ranges from $d = 0.02$ to $d = 0.21$ (Bem et al., 2016). These two values describe a ratio of 1 : 9.7, indicating that the observed object-level effects are very heterogeneous. The heterogeneity of the observed object-level effects may suggest that the *average* psi-effect should be evaluated by combining a statistical inference strategy with an error account (Lord and Novick, 1968). This evaluation, however, would remain sensitive to how n and s vary across individual studies. But as statistical parameters, n and s lack theoretical meaning. Particularly s is merely a normalization factor to render several object-level effects comparable.

To achieve an evaluation that is independent of how n and s vary across the object-level studies, one should rather compare the point-specific ES_{THEO} *directly* to the point-specific ES_{OBS} in each study, *without* averaging the effect. To this end, Bem et al.'s (2016; Supplementary Table A1) meta-analytical findings can be re-analyze as follows:

1. As Bem himself proposed (Bem, 2011, p. 409, note 1), the theoretical psi-effect is specified as $d_{THEO} = 0.20$ using a scale of z-values where $s = 1$. Consequently, $d_{THEO} = ES_{THEO}$. (A theoretical construct cannot reasonably predict a smaller psi-effect because it would be overlain by the standard measurement error.)
2. To control for the quality of the object-level studies, we exclude the 49 non-peer-reviewed object-level studies, retaining the 51 peer-reviewed ones (see Bem et al., 2016, Supplementary Table S1).
3. To eliminate the variation of s, the mean difference $(m_1 - m_0)$ is calculated by multiplying the instance of ES_{OBS} in each peer-reviewed object-level study with that study's observed s. This yields $ES_{OBS} = (m_1 - m_0) / s$, where $s = 1$.
4. For each peer-reviewed object-level study, I_{SIM} is computed as follows: (a) $I_{SIM} = 0$ if the mean difference is negative; (b) I_{SIM} is undefined if the between-group ES_{OBS}-difference (treatment vs. control) is 0; otherwise, since $s = 1$, (c) $I_{SIM} = (ES_{THEO} = 0.20 \times s) / (ES_{OBS} \times s) = (0.20 / ES_{OBS})$.

Because s has been eliminated, the 95%-SI [0.80;1.20] can be applied to each peer-reviewed object-level study individually. The two relevant parameters are $ES_{THEO} = d_{THEO} = 0.20$ relative to the sample size of an object-level study, and the percentage of ES_{OBS}-instances that fall inside the 95%-SI given $ES_{THEO} = d_{THEO} = 0.20$.

The application of I_{SIM} indicates that, although each of the 51 peer-reviewed object-level studies was published as evidence *for* a psi-effect (Bem et al., 2016), the mean difference is negative ($I_{SIM} = 0$) in 16 studies (31% of 51 studies), that two studies show no difference (I_{SIM} is undefined), and that the I_{SIM}-value falls outside the 95%-SI in 22 studies (43%). This means that $\mathrm{ES}_{\mathrm{OBS}}$ is *insufficiently similar* to $\mathrm{ES}_{\mathrm{THEO}}$.

In the remaining 11 studies (22% of 51 studies), where $\mathrm{ES}_{\mathrm{OBS}}$ is *sufficiently similar* to $\mathrm{ES}_{\mathrm{THEO}}$, the percentages of I_{SIM}-values falling inside the 95%-SI (see Table 1) are nevertheless quite low: 33% ($n_0 = n_1 = 100$); 37% (150); 33% (99); 33% (100); 33% (100); 34% (109); 23% (49); 33% (100); 34% (111); 42% (201); 23% (50). This means that each study's sample is *too small* to generate the evidence required to consider empirically adequate a theoretical construct that predicts $\mathrm{ES}_{\mathrm{THEO}} = d_{\mathrm{THEO}} = 0.20$.

To appreciate the sample size that is needed to consider as empirically adequate a theoretical construct that predicts $\mathrm{ES}_{\mathrm{THEO}} = d_{\mathrm{THEO}} = 0.20$, a one-sided t-test under $\alpha = 0.05$ and test-power of $(1 - \beta) = 0.80$ already requires $n_0 = n_1 = 101$. Under $\alpha = \beta = 0.05$, it even requires $n_0 = n_1 = 201$. The reason for the large samples is that the theoretically predicted effect is small enough to be accounted for exclusively by random influences on the empirical setting. But random influences are independent of $\mathrm{ES}_{\mathrm{THEO}}$ and so lack theoretical meaning. Indeed, this is the reason why $\mathrm{ES}_{\mathrm{THEO}} = d_{\mathrm{THEO}} = 0.20$ requires a statistical corroboration against random influences in the first place.

In sum, although $\mathrm{ES}_{\mathrm{OBS}}$ is sufficiently similar to $\mathrm{ES}_{\mathrm{THEO}}$ in 11 out of 51 peer-reviewed object-level studies, these 11 studies *individually* fail to provide the evidence required to consider as empirically adequate a theoretical construct that predicts $\mathrm{ES}_{\mathrm{THEO}} = d_{\mathrm{THEO}} = 0.20$. Arguably, therefore, if the empirical adequacy of the theoretically predicted psi-effect had been evaluated before conducting additional studies, some research effort concerning the psi-effect could have been avoided.

Discussion

Whether a theoretical construct adequately predicts future observations is a distinct question from whether a data-based parameter estimate (induced from past observations) deviates statistically significantly from a random distribution. This difference matters because behavioral science research regularly uses a data-based parameter estimate and its associated confidence bounds as a proxy for a theoretical construct. But a parameter that is estimated using a z-standardized effect size measure such as Cohen's d cannot distinguish whether particularly a *small* observed d-value points to a mean difference that is too small to be observable, or rather to a large s. Without *making* this distinction, however, the evaluation of the empirical adequacy of a theoretical construct is out of reach.

The I_{SIM} measure and the SI fare better. Both together can inform the evaluation of the empirical adequacy of a theoretical construct because, if the inductive element s that serves to z-standardize measurements is avoided, then the observed mean difference *ceases* to be "blurred" by random influences. As this enables a *direct* comparison between the theoretically predicted and the observed mean-difference, the evaluation of the empirical adequacy of a theoretical construct is placed within reach. On how I_{SIM} and the SI can be applied beyond a simple experimental setting, see Supplementary Appendix S1. For additional examples, see Supplementary Appendix S2. To apply I_{SIM} and the SI to extant data, we provide an online tool at https://adrian-stanciu.shinyapps.io/Similarity-Index/.

Practical implications

As behavioral science has come under scrutiny, *replication crisis* denotes that few previously "established" findings are independently replicable and that questionable research practices are regularly employed (e.g., Kerr, 1998; Klein, 2014; Irvine, 2021; Nosek et al., 2022). A familiar response to the replication crisis is to recommend measures that improve the quality of data (e.g., study pre-registrations, multi-lab projects, or open access to materials). Such measures constitute important elements of an inductive approach to parameter estimation. But some effort must also go toward developing theoretical constructs that logically entail an empirically adequate prediction, i.e., toward a deductive approach to theory construction.

A central limitation of the inductive approach to parameter estimation is exemplified by meta-analytical research. To arrive at robust meta-level or population effect size estimates, observed object-level effects are regularly sought to be made comparable by weighing them to the observed s (Schulze, 2004). But since s varies with the (random) influences on an empirical setting, this *invites* all the problems discussed above. So, if a meta-analysis retains the observed s of observed object-level effects, a robust meta-level or population effect size estimate cannot be had. For this reason, s should be avoided in both theory construction research and meta-analytical research.

The similarity index I_{SIM} fares better. First, I_{SIM} offers a more transparent view of observations. This can assist in improving a theoretical construct because using I_{SIM} and the associated 95%-SI allows distinguishing between an empirically adequate prediction (true positive; $\mathrm{ES}_{\mathrm{THEO}} = \mathrm{ES}_{\mathrm{POP}}$) and an empirically inadequate one (false positive; $\mathrm{ES}_{\mathrm{THEO}} \neq \mathrm{ES}_{\mathrm{POP}}$). Making this distinction is required to decide whether a theoretical construct can be maintained, whether its theoretically predicted effect should be adjusted, or whether additional data should be collected. The last option particularly counts if available data indicate a small effect, which is generally not well-observable.

Second, the I_{SIM} measure and the 95%-SI help to evaluate whether a false positive prediction indicates that the population effect is under- or overestimated. After all, for all possible combinations of a theoretically predicted effect and a sample size, as long as the percentage of non-matching observations ($\mathrm{ES}_{\mathrm{THEO}} \neq \mathrm{ES}_{\mathrm{POP}}$) makes it unreasonable to evaluate the $\mathrm{ES}_{\mathrm{THEO}}$-value as a true positive prediction, an empirically adequate prediction is more probable to fall inside the 95%-SI than not.

Third, assume that, as n increases, also the value of $\mathrm{ES}_{\mathrm{OBS}}$ becomes increasingly more similar to the value of the true population parameter (law of large numbers). If so, then the corresponding increase in the percentage difference between a true positive and a false positive prediction goes along with an increase in the proportion of viable theoretical assumptions relative to all possible alternative theoretical assumptions. With each additional I_{SIM}-value for a point-$ES_{THEO} = x$ that falls inside the 95%-SI, therefore, it becomes more reasonable for researchers to develop a theoretical construct for *x* because "getting something right" about *x* is more probable than not.

Fourth, if additional *independent* studies happen to estimate a point-$ES_{OBS} = y$ that is similar to *x*, then I_{SIM} continues to approximate the condition for a perfect match between prediction and observations ($I_{SIM} = 1$). The independence of additional studies entails that the approximation of $I_{SIM} = 1$ is unlikely to occur by random. Consequently, a researcher's confidence that $ES_{THEO} = x$ is empirically adequate would increase. The same rationale underlies having confidence in a meta-analytically estimated point-ES_{OBS} that is based on independently observed object-level effects (Hunter and Schmidt, 2004).

The use-value of an I_{SIM}-based evaluation of a theoretical construct is perhaps most readily apparent in the context of the *research program strategy* (RPS) (Witte and Zenker, 2017; Krefeld-Schwalb et al., 2018). If the effects of several independent and topically related studies are observed under low error-rates, then RPS induces the observed mean effect as a parameter estimate (see the subsection *Parameter Estimation*). Next, RPS develops a theoretical construct that logically entails a theoretically predicted point-effect of identical size as this inductive parameter estimate. Provided *new* observations under low error-rates, finally, if the likelihood of the theoretically predicted effect sufficiently exceeds the likelihood of an alternative effect, then RPS evaluates the former as *preliminarily* verified, respectively as *substantially* verified if the likelihood of the theoretically predicted effect is sufficiently similar to the maximum likelihood of new observations. For the verification thresholds of this statistical likelihood model, see Krefeld-Schwalb et al. (2018, p. 22).

Beyond this likelihood model, the attempt to verify a theoretically predicted effect by comparing it to observations *requires* an I_{SIM}-like measure. An inductive parameter estimate, after all, has uncertainty bounds that reflect the variance of observations, whereas a theoretical construct that is developed based on theoretical considerations predicts a point-specific effect. For this reason, I_{SIM} avoids comparing the theoretically predicted effect *indirectly* to observations, an indirectness that results from using a statistical error account and a data distribution (e.g., a *t*-, *F*-, or X^2-distribution). Instead, the theoretically predicted effect is compared *directly* to observations (as measured), while the admissible variation of a theoretical construct is captured by the 95%-SI (see the section "*Case study*").

This explains why we modeled the admissible variation of a theoretical construct by simulating random samples of possible measurements, rather than by using an inferential statistical theory (e.g., a likelihood model). In RPS, the inferential statistical evaluation of (simulated or real) observations is useful, only if the I_{SIM}-value already lies within the 95%-SI, indicating that the theoretically predicted effect is similar to observations. Thus, I_{SIM} evaluates the similarity between a theoretical construct and observations *before* inferentially testing the theoretically predicted effect (Witte and Heitkamp, 2006). Nevertheless, for a specific theoretically predicted effect to be accepted as empirically adequate, both its point-specification and its statistical substantial verification are required. In brief, I_{SIM} assists in specifying the effect size, while RPS verifies it.

Limitations

Rather than replacing standardized effect size measures such as Cohen's *d* or inductive data-evaluation tools like a model-fitting index, I_{SIM} complements them. I_{SIM} should be applied mindfully. Several limitations apply:

First, I_{SIM} does not offer a criterion for a data-based decision to accept or reject hypotheses. Rather than comparing two hypotheses (H_0, H_1) in view of data, I_{SIM} evaluates only the H_1-hypothesis that states ES_{THEO}. Therefore, I_{SIM} cannot enable a relative statistical corroboration of a theoretical construct against random influences. This continues to require statistical testing.

Second, if the theoretically predicted effect $ES_{THEO} = x$ falls outside the 95%-SI, then *x* appears to be empirically *inadequate*. This appearance may mislead researchers to prematurely abandon *x* as a candidate value for ES_{THEO}. But as a rule, the decision to abandon *x* should squarely depend on having collected an adequately large sample.

Third, like all formal measures, I_{SIM} is open to "tweaking" the data to let ES_{OBS} and ES_{THEO} match artificially. With a new formal measure, therefore, additional temptation to engage in questionable research practices may arise.

Fourth, a simple "recycling" of the ES_{OBS}-value as the ES_{THEO}-value would *trivially* satisfy the perfect-match condition ($I_{SIM} = 1$), known as *p*-harking. So, the same critical considerations apply as were stated immediately above (Kerr, 1998).

Fifth, in the context of a confirmatory factor analysis (CFA), which relies on an explorative factor analysis (EFA) to evaluate the deviation of predetermined parameters in some complex mathematical model, several of these parameters must be determined simultaneously (e.g., the number and correlations of factors, their weights, loadings, etc.). However, I_{SIM} cannot be applied to test whether the complex mathematical model itself agrees with the abstract data deduced from it; I_{SIM} can only test whether a basic parameter (e.g., a mean or a correlation) agrees with empirical data. Given a correlation matrix, for instance, I_{SIM} can evaluate the similarity between a single predicted correlation and an empirically observed correlation (see Supplementary Appendix S2, personality traits and life outcomes). As a basic (non-complex) measure, I_{SIM} thus operates at the level of each element in a correlation matrix and can there compare a prediction *directly* with observations (see Perez-Gil et al., 2000).

Conclusion

The identification of an empirically adequate theoretical construct requires determining whether a theoretically predicted effect is sufficiently similar to an observed effect. To this end, we proposed I_{SIM} and the 95%-SI as a simple measure to evaluate the similarity between a theoretically predicted effect and observations, a measure that avoids the statistical element of the observed standard deviation. Using computer simulations, we estimated the sample size and the observed effect size that are necessary to identify an empirically adequate theoretical construct.

Generally relevant for theory construction research, the I_{SIM} measure and the 95%-SI particularly serve to develop a point-specific theoretical construct, where both should be applied alongside a statistical corroboration measure (e.g., the likelihood ratio). If the I_{SIM}-value falls within the 95%-SI, then a theoretical construct postulating a theoretically predicted point-specific effect $\text{ES}_{\text{THEO}} = x$ can be (fallibly) maintained as empirically adequate. If independent studies subsequently observe a point-effect $\text{ES}_{\text{OBS}} = y$ that is similar to x, a researcher's confidence that x is empirically adequate would increase. Whereas if too many I_{SIM}-values fall outside the 95%-SI as the number of independent studies increases, then $\text{ES}_{\text{THEO}} = x$ must be corrected, or the standard error must be reduced, e.g., by restricting the experimental setting. The most direct way of reducing the standard error, of course, is to increase the sample.

An exemplary application of I_{SIM} to recent meta-analytical findings on the precognition effect (Bem et al., 2016) indicated that 51 peer-reviewed object-level studies individually fail to provide the evidence that is required to evaluate as empirically adequate a theoretical construct that predicts a precognition effect of d =0.20 (additional application examples are found in Supplementary Appendix S2).

In behavioral science as elsewhere, measurement comprises an ontological aspect related to the theoretical construct under development, and an epistemological aspect related to the specific measurement procedures employed. When using Cohen's d measure, behavioral scientists tend to address a question that combines both aspects of measurement. This is understandable if theory-testing relies on statistical inference procedures, which simultaneously relate to both aspects of measurement. But to facilitate theory construction research and the development of measurement, the ontological and epistemological aspects are best kept separate. Otherwise, it is quite difficult to say what a measurement instance in fact measures.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/rgwsp/.

Author contributions

EHW developed the I_{SIM} measure. AS coded and ran the simulations. All authors drafted the manuscript, which FZ edited. All authors approved the final submitted version.

Funding

FZ acknowledges support from TUBITAK (No. 118C257).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.980261/full#supplementary-material

References

Andreas, H. (2021). "Theoretical Terms in Science," in *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*. ed. E. N. Edward Available at: https://plato.stanford.edu/archives/fall2021/entries/theoretical-terms-science/

Bem, D. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J. Pers. Soc. Psychol.* 100, 407–425. doi: 10.1037/a0021524

Bem, D., Tressoldi, P., Rabeyron, T., and Duggan, M. (2016). Feeling the future: a meta-analysis of 90 experiments on the anticipation of random future events [version 2; referees: 2 approved]. F1000 research. 4:1188. https://f1000researchdata.s3.amazonaws.com/datasets/7177/9efe17e0-4b70-4f10-9945-a309e42de2c4_TableA1.xlsx]

Bollen, K. A., Bauer, D. J., Christ, S. L., and Edwards, M. C. (2010). "An overview of structural equations models and recent extensions" in *Recent developments in* *social science statistics*. eds. S. Kolenikov, D. Steinley and L. Thombs (Hoboken: Wiley), 37–80.

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33, 261–304.

Cardena, E. (2018). The experimental evidence for parapsychological phenomena: a review. *Am. Psychol.* 73, 663–677. doi: 10.1037/amp0000236

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cornelissen, J., Höllerer, M. A., and Seidl, D. (2021). What theory is and can be: forms of theorizing in organizational scholarship. *Organ. Theory* 2, 263178772110203–263178772110219.

Eronen, M. I., and Bringmann, L. F. (2021). The theory crisis in psychology: how to move forward. *Perspect. Psychol. Sci.* 16, 779–788. doi: 10.1177/1745691620970586

Eronen, M. I., and Romeijn, J. W. (2020). Philosophy of science and the formalization of psychological theory. *Theory Psychol.* 30, 786–799.

Fiedler, K., and Prager, J. (2018). The regression trap and other pitfalls of replication science—illustrated by the report of the Open Science collaboration. *Basic Appl. Soc. Psychol.* 40, 115–124. doi: 10.1080/01973533.2017.1421953

Fleck, L. (1935). *Genesis and development of a scientific fact*. Chicago: The University of Chicago Press.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personal. Soc. Psychol. Bull.* 44, 16–23.

Gelman, A., and Carlin, J. (2014). Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect. Psychol. Sci.* 9, 641–651.

Gervais, W. M. (2021). Practical methodological reform needs good theory. *Perspect. Psychol. Sci.* 16, 827–843. doi: 10.1177/1745691620977471

Gigerenzer, G. (1998). Surrogates for theories. *Theory Psychol.* 8, 195–204.

Hempel, C. G. (1988). Provisoes: a problem concerning the inferential function of scientific theories. *Erkenntnis* 28, 147–164. doi: 10.1007/BF00166441

Henderson, L. (2020). The problem of induction. in E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (online book)*. Stanford: Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/entries/induction-problem/

Hume, D. (1739). *A treatise of human nature*. Oxford: Oxford University Press.

Hunter, J. E., and Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. *2nd* Edn. Thousand Oaks: Sage Publications.

Irvine, E. (2021). The role of replication studies in theory building. *Perspect. Psychol. Sci.* 16, 844–853. doi: 10.1177/1745691620970558

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* 2, 196–217.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory Psychol.* 24, 326–338.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B. Jr., Alper, S., et al. (2018). Many labs 2: investigating variation in replicability across sample and setting. *Adv. Methods Pract. Psychol. Sci.* 1, 443–490.

Krefeld-Schwalb, A., Witte, E. H., and Zenker, F. (2018). Hypothesis-testing demands trustworthy data—a simulation approach to statistical inference advocating the research program strategy. *Front. Psychol.* 9:460. doi: 10.3389/fpsyg.2018.00460

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: The University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research Programmes*. Cambridge: Cambridge University Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* 4:863. doi: 10.3389/fpsyg.2013.00863

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence testing for psychological research: a tutorial. *Adv. Methods Pract. Psychol. Sci.* 1, 259–269.

Linden, A. H., and Hönekopp, J. (2021). Heterogeneity of research results: a new perspective from which to assess and promote progress in psychological science. *Perspect. Psychol. Sci.* 16, 358–376. doi: 10.1177/1745691620964193

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: sir Karl, sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46, 806–834. doi: 10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant it. *Psychol. Inq.* 1, 108–141. doi: 10.1207/s15327965pli0102_1

Meehl, P. E. (1992). Cliometric metatheory: the actuarial approach to empirical, history-based philosophy of science. *Psychol. Rep.* 91, 339–404. doi: 10.2466/pr0.2002.91.2.339

Meehl, P. E. (1997). "The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numeral predictions" in *What if there were no significance tests?* eds. L. L. Harlow, S. A. Mulaik and J. H. Steiger (Mahwah: Erlbaum), 393–425.

Miłkowski, M., Hohol, M., and Nowakowski, P. (2019). Mechanisms in psychology: the road towards unity? *Theory Psychol.* 29, 567–578.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* 38, 2074–2102. doi: 10.1002/sim.8086

Muthukrishna, M., and Henrich, J. (2019). A problem in theory. *Nat. Hum. Behav.* 3, 221–229. doi: 10.1038/s41562-018-0522-1

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology* 44, 190–204.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* 73, 719–748. doi: 10.1146/annurev-psych-020821-114157

Oberauer, K., and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* 26, 1596–1618. doi: 10.3758/s13423-019-01645-2

Olsson-Collentine, A., Wicherts, J. M., and van Assen, M. A. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychol. Bull.* 146, 922–940. doi: 10.1037/bul0000294

Peirce, C. S. (1931–1958). in *Collected papers of Charles Sanders Peirce*. eds. P. Weiss, C. Hartshorne and A. W. Burks, vol. *1–8*. Cambridge, MA: Harvard University Press.

Perez-Gil, J. A., Moscoso, S. C., and Rodriguez, R. M. (2000). Validez de constructo: el uso de analisis factorial exploratorio-confirmatorio Para obtener evidencias de validez [construct validity: the use of exploratory-confirmatory factor analysis in determining validity evidence]. *Psicothema* 12, 442–446.

Popper, K. R. (1959). *Logic of discovery*. London: Routledge.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Available at: https://www.R-project.org/

Schäfer, T., and Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases. *Front. Psychol.* 10:813. doi: 10.3389/fpsyg.2019.00813

Schauer, J. M., and Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychol. Bull.* 146, 701–719. doi: 10.1037/bul0000232

Schulze, R. (2004). *Meta-Analysis*. A comparison of approaches. Cambridge: Hogrefe & Huber.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2013). Life after p-hacking. *Meet. Soc. Pers. Soc. Psychol.* doi: 10.2139/ssrn.2205186

Szucs, D., and Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* 15:e2000797. doi: 10.1371/journal.pbio.2000797

Torchiano, M. (2020). Effsize: efficient effect size computation (package version 0.8.1.). doi: 10.5281/zenodo.1480624,

van Fraassen, B. (1980). *The scientific image*. Oxford: Oxford University Press. doi: 10.1093/0198244274.001.0001

van Rooij, I., and Baggio, G. (2020). Theory before the test: how to build high-verisimilitude explanatory theories in psychological science. *Perspect. Psychol. Sci.* 16, 682–697. doi: 10.1177/1745691620970604

Wagenmakers, E., and Farrell, S. (2004). AIC model selection using Akaike weights. *Psychon. Bull. Rev.* 11, 192–196. doi: 10.3758/BF03206482

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., Francois, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686

Wickham, H., Francois, R., Henry, L., and Müller, K. (2021). Dplyr: a grammar of data manipulation (package version 1.0.7.). Available at: https://CRAN.R-project.org/package=dplyr

Witte, E. H., and Heitkamp, I. (2006). Quantitative Rekonstruktionen (Retrognosen) als Instrument der Theorienbildung und Theorienprüfung in der Sozialpsychologie. *Z. Sozialpsychol.* 37, 205–214. doi: 10.1024/0044-3514.37.3.205

Witte, E. H., and Zenker, F. (2017). From discovery to justification: outline of an ideal research program in empirical psychology. *Front. Psychol.* 8:1847. doi: 10.3389/fpsyg.2017.01847

frontiers | Frontiers in Psychology

TYPE Hypothesis and Theory
PUBLISHED 09 December 2022
DOI 10.3389/fpsyg.2022.933183

Awareness: An empirical model

Federico Bizzarri[1], Alessandro Giuliani[2] and Chiara Mocenni[1]*

[1]Department of Information Engineering and Mathematics, University of Siena, Siena, Italy,
[2]Environment and Health Department, Istituto Superiore di Sanità, Rome, Italy

Check for updates

OPEN ACCESS

EDITED BY
Edoardo Datteri,
University of Milano-Bicocca, Italy

REVIEWED BY
Athanasios Drigas,
National Centre of Scientific Research Demokritos, Greece
Giulio de Felice,
Sapienza University of Rome, Italy
Jianwei Shen,
North China University of Water Resources and Electric Power, China

*CORRESPONDENCE
Chiara Mocenni
chiara.mocenni@unisi.it

SPECIALTY SECTION
This article was submitted to Theoretical and Philosophical Psychology,
a section of the journal Frontiers in Psychology

RECEIVED 30 April 2022
ACCEPTED 21 November 2022
PUBLISHED 09 December 2022

CITATION
Bizzarri F, Giuliani A and Mocenni C (2022) Awareness: An empirical model. *Front. Psychol.* 13:933183.
doi: 10.3389/fpsyg.2022.933183

COPYRIGHT
© 2022 Bizzarri, Giuliani and Mocenni. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

In this work, we face the time-honored problem of the contraposition/integration of analytical and intuitive knowledge, and the impact of such interconnection on the onset of awareness resulting from human decision-making processes. Borrowing the definitions of concepts like intuition, tacit knowledge, uncertainty, metacognition, and emotions from the philosophical, psychological, decision theory, and economic points of view, we propose a skeletonized mathematical model grounded on Markov Decision Processes of these multifaceted concepts. Behavioral patterns that emerged from the solutions of the model enabled us to understand some relevant properties of the interaction between explicit (mainly analytical) and implicit (mainly holistic) knowledge. The impact of the roles played by the same factors for both styles of reasoning and different stages of the decision process has been evaluated. We have found that awareness emerges as a dynamic process allowing the decision-maker to switch from habitual to optimal behavior, resulting from a feedback mechanism of self-observation. Furthermore, emotions are embedded in the model as inner factors, possibly fostering the onset of awareness.

KEYWORDS

tacit knowledge, cognition, decision making, uncertainty, Markov models, machine intelligence, optimization, intuition

Introduction

Aside from the classical analytical perspective, this work mainly focuses on the weight of the impact and relevance of other facets that belong to the decision-making processes, such as tacit knowledge, intuition, emotions, awareness, and self-awareness. These factors, once considered largely irrelevant (if not a nuisance to be eliminated) in the decision-making process, are now being taken into account more seriously in cognitive studies, and their multi-faceted effects are intensely analyzed. Even though all of them have been thoroughly studied and described from a theoretical point of view in different fields of investigation, a formalization from a modeling point of view is still missing. This is the gap the present work aspires to begin filling by proposing a possible modeling formulation that is both broad enough to consider all these different aspects and sufficiently simple to be clearly understandable, and thus, interpretable according to the different perspectives of the practitioner. Although limited and imperfect by nature, also due to the difficulties in modeling complex phenomena—as human decisions are—this study could contribute to introducing new aspects which can expand research in the field of decision-making.

This research has a multidisciplinary intent, involving different disciplines ranging from behavioral and cognitive sciences, economy, mathematics, philosophy, and psychology, and attempts to incorporate all the different perspectives involved in the process. We tried to be as faithful as possible to the different concepts exposed according to specialist literature while maintaining a constant interest and focus on the modeling perspective. Indeed, this work does not intend to give unique and definitive modeling recipes; on the contrary is aimed at fostering general interest in the explicit inclusion of crucial aspects of decision-making into quantitative models of awareness. Specifically, our aim is to propose a mathematical model incorporating all processes fostering the emergence of awareness. On one hand by identifying the main characteristics behind decision-making (including analytic and intuitive factors), their relationship with emotions and other drivers, and on the other hand by highlighting novel logical and philosophical aspects such as the importance of tacit knowledge, the correlation between optimal decisions and uncertainty, and the awareness dynamics. The mathematical formulation of the model is grounded on Decision Markov Processes, which embed most characteristics of human decision-making, indeed they integrate control actions, uncertainties, and temporal dynamics. Specifically, at each time step the decision-maker's specific choice provokes a change in the system's state according to the control actions he/she chooses. Additionally, this change is affected by external noisy stimuli. The evolutionary dynamics of such a system can be evaluated for very long-time processes (up to the limit of infinite time). For finite time-horizon processes, such as the ones reasonably considered in this paper, the trajectory of the system depends on the reward functions, the transition probabilities, and, at the final time, the terminal reward value. In any case, the convergence of the trajectory will be evaluated on average due to the stochastic nature of the process.

The formulation of mathematical models enables us to perform extensive simulations to understand the multifaceted nature of this process.

The paper is organized as follows. Section Decision-making processes: an overview exposes general concepts related to decision-making processes and awareness, as the individual styles and approaches, the role of emotions, tacit knowledge, and the relationship between decisions and information. Furthermore, we will offer a panoramic view on how scientific literature deals with the concept of awareness and its impact on decisions. Section Developing a mathematical model of awareness reports the mathematical formulation of the proposed model. It is first simply and generically described, and later possible developments and extensions are introduced. Section Numerical results presents some numerical results of the simulations carried out by applying the proposed model and their discussion.

Decision-making processes: An overview

Rationality, intuition, tacit knowledge, and emotions

Rationality and intuition

Until very recently, scholars and practitioners agreed that effective decision-making occurs only under the most rational conditions. Since Descartes (2008) cognition has had a stronghold as being the only legitimized contributor to reasoning in that decisions must come exclusively from rational, cognitive, and logical processes, while emotions, intuition, and other subjective aspects are not considered as having a significant role in the process. This conventional teaching has been that "the more objective and rigorous our thinking processes are, the better our decisions will be." A totally different perspective (much more similar to contemporaneous views) on cognition was outlined in the same years by another French scientist and philosopher, Blaise Pascal (2012). It is worth reporting a small extract of his considerations on cognition processes, which are at the very basis of our proposal (emphasis added).

> *"THE DIFFERENCE between the mathematical and the intuitive mind.—In the one, the principles are palpable, but* REMOVED FROM ORDINARY USE; *so that for want of habit it is difficult to turn one's mind in that direction: but if one turns it thither ever so little, one sees the principles fully, and one must have a quite inaccurate mind who reasons wrongly from principles so plain that it is almost impossible they should escape notice [...] But in the intuitive mind,* THE PRINCIPLES ARE FOUND IN COMMON USE *and are before the eyes of everybody. One has only to look, and no effort is necessary; it is only a question of* GOOD EYESIGHT, *but it must be good, for the principles are so subtle and so numerous, that it is almost impossible but that some escape notice [...] These principles are so fine and so numerous that a very delicate and very clear sense is needed to perceive them and to judge rightly and justly when they are perceived,* WITHOUT FOR THE MOST PART BEING ABLE TO DEMONSTRATE THEM *in order as in mathematics;* BECAUSE THE PRINCIPLES ARE NOT KNOWN TO US IN THE SAME WAY, AND BECAUSE IT WOULD BE AN ENDLESS MATTER TO UNDERTAKE IT. WE MUST SEE THE MATTER AT ONCE, AT ONE GLANCE, *and not by a process of reasoning, at least to a certain degree [...]".*

Based on the above statements, it is clear that Pascal considered "intuition" ("*esprit de finesse*" in French, as opposed to the "*esprit de géométrie*") a proper form of knowledge and not "irrational" and "purely emotive" nuisances to the proper way of reasoning. In another part of his essay, he clearly states that a real scientist (philosopher) must adopt both attitudes to have a fruitful approach to science. In real life we have never had problems accepting Pascal's view, and it is shared common sense to appreciate both the "holistic-intuitive" and "logical-analytic" capacity of decision-makers (whether they are managers, scientists, physicians, etc.). However, shifting from real-life to academic formalism, things abruptly change. The Cartesian way of thinking, much more primitive with respect to Pascal, has been reinforced in more recent times with the ascendancy of empiricism and positivism (Tayor, 2004).

Given the gaps in the rational theories, an alternative perspective proposed by theorists calls for a richer conception of the decision-maker accounting for the assumption that decisions are also driven by emotion, intuition, imagination, experience, and memories, and thus implicitly reviving Pascal's statements. There is now a consistent body of research delving into the nature of decision-making, particularly into the role of cognition, intuition, and emotion in human decisions (Soosalu et al., 2019). Notably, intuition and cognition deal with two different ways to process information, which we call *intuitive* and *analytical* (Adinolfi and Loia, 2021). Although dual-process theories come in several forms, they reflect the generic fundamental distinction between the two processes. The first is related to intuition, which is often considered relatively undemanding in terms of cognitive resources, and is associative, tacit, intuitive, and holistic; hence, at odds with Pascal's position as he considers intuitive and analytical knowledge to be of equal importance. On the contrary, the second involves conscious, analytical, deliberate, cognitive, logical, linear, and reason-oriented thinking, making certain demands on "cognitive" resources (Hodgkinson et al., 2008; Kahneman, 2017). In our opinion, this consideration holds true if (and only if) we consider such resources in terms of computational cost and time, but not in terms of depth and subtlety. It is no accident that the English translation of "esprit de finesse" as "intuition" conveys a somewhat different meaning. The etymological roots of the term *intuition* stem from the Latin word *in-tuir*, which can be translated as "looking, regarding or knowing from within". Intuition encompasses a complex set of interrelated cognitive, affective, and somatic processes in which there is no apparent intrusion of deliberate, rational thought. Moreover, the outcome of this process (an intuition) occurs almost instantaneously and can be difficult to articulate. The outcomes of intuition are perceived as a holistic "hunch", a sense of calling or overpowering certainty, and an awareness of knowledge that is on the threshold of conscious perception (Bechara and Damasio, 2005). In their comprehensive review of literature on intuition within the field of management, Dane and Pratt defined intuition as "*affectively charged judgment that arises through rapid, non-conscious, and holistic associations*" (Dane, 2007; Adinolfi and Loia, 2021). We can say that Blaise Pascal's cognition theory has obtained its deserved consideration after nearly 350 years.

Tacit knowledge

The roles of intuition and tacit knowledge were formally incorporated into the theory (Brockmann and Anthony, 1998) as factors that lead to better decisions compared to those relying solely on the rational or analytical approach. This is particularly evident in those areas involving creativity such as innovating, visioning, and planning. The concept of *tacit knowledge* (sometimes also called *implicit knowledge*) is mainly attributed to Michael Polanyi, who introduced it for the first time in 1958 in his work Personal Knowledge (Polanyi, 2009). This term indicates a kind of knowledge that is opposed to formal and codified explicit knowledge; it is difficult to express or extract and thus even more difficult to transfer to others through writing or speech (Polanyi, 2009). It represents a content that is neither part of one's normal consciousness nor open to introspection. When applied, tacit knowledge is helpful but not externally expressed or declared; rarely do we recognize when we are using tacit knowledge. According to Polanyi, people cannot describe their use of tacit knowledge; "*we simply know more than we can tell*" is his typical expression to explain this concept. Implicit learning and implicit knowledge contribute to the knowledge structures upon which individuals draw when making intuitive judgments (see the visual representation reported in Figure 1). However, although they may underpin the non-conscious cognitive, affective, and somatic processes that lead to an intuitive judgment, they are not equivalent to intuition (Reber, 1993; Hodgkinson et al., 2008). In our opinion, tacit knowledge is the closest relative to Pascal's "*esprit de finesse*" which, in its original definition, is devoid of any "emotional" aspect.

Emotions

As Damasio points out, an important aspect of the purely rational position is that to obtain the best results we must keep emotions out (Damasio, 1994). For a long time, emotion has been largely banished from the predominant philosophies and theories regarding decision, reason, and management. However, emotions have a considerable impact on an individual's decisions and must be carefully considered, particularly the *immediate emotions*. Immediate emotions are those experienced at the moment of decision, in contrast to the ones expected to be experienced in the future, like regret and disappointment (Loewenstein, 2000).

We most certainly always mix emotion and reasoning, analytic attitude and intuition, this mixture being something neuroscientists consider essential (Damasio, 2012). Rational decision-making skills are required to clearly and logically process available information, and thus allow for accurate

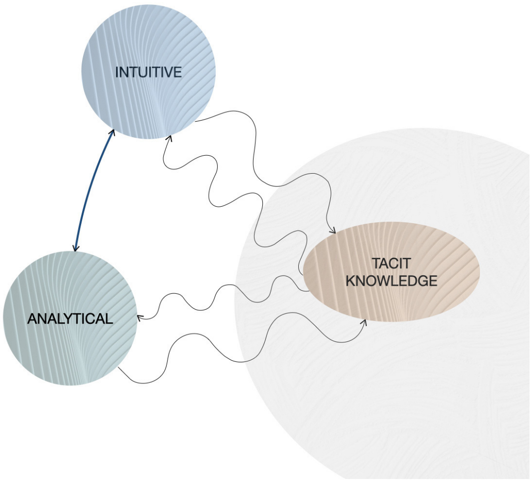


FIGURE 1
A schematic representation of Knowledge. Tacit knowledge is involved in both intuitive and analytical reasoning, yet it is more heavily used in the former while in the latter it is only slightly adopted. An example of the application of tacit knowledge in analytical reasoning could come from the field of data analysis, where we can recognize the action of tacit knowledge in the choice of the model to use to analyze collected data. In both cases it is a *fast* process: tacit knowledge is immediately available and instantaneously applied by the individual. On the other hand, both intuitive and analytical approaches, thanks to the experience they bring to the individual, contribute to the *slow* process of knowledge sedimentation that creates an individual's unique implicit knowledge. These aspects are represented in the image by links: the *fast* (*slow*) characteristic is figuratively represented by waves with a high (low) frequency. Their amplitude represents the weight of the influence between tacit knowledge and analytical/intuitive reasoning and, on the contrary, the influence of analytical/intuitive approaches on tacit knowledge. Moreover, the characteristic of tacit knowledge that can be neither explicitly declared nor open to introspection is represented by the haziness of its area. The last arrow, between analytical and intuitive, stands for the continuously evolving relationship between these two individual modalities, which has a speed of change entirely depending on individual characteristics.

perception and interpretation of events. In addition to this type of knowledge it is essential to consider that people make decisions based on tacit knowledge grounded in experience, and may use intuitive decision strategies almost exclusively, particularly under high-stress conditions (Sayegh et al., 2004). All these aspects are, to some extent, included in the model proposed in this article, which integrates the analytical, logical, and cognitive abilities of the decision-maker but also subjective aspects like intuition, tacit knowledge, and emotions. These components are always present in any decision-making process and must be collectively considered to reach more aware choices, which can, in turn, lead to an individual's all-encompassing well-being.

Information and overfitting

A crucial point in the decision-making process is how a single individual approaches the incoming information. We could notice how decisions become faster and faster and are made in a constantly changing environment. The amount of data grows exponentially, but despite its abundance it could be inaccurate, incomplete, or confusing. The consequent increasing spread of misinformation is more serious in some sectors with respect to others: a paradigmatic case is the increasing interest in the field of *Infodemiology*, the study of the determinants and distribution of health information and misinformation (Eysenbach, 2002). This phenomenon represents the importance in exploring the correlation between data, reasoning propensity, and decisions. When we think about thinking it is easy to assume that "more is better"; we may assume that having more data about the decision context can lead to a more accurate prediction of the future consequences of our choices, and thus to a better result. However, this cannot be always true. The question of how hard to think and how many factors to consider is at the heart of a thorny problem that statisticians and machine-learning researchers call *Overfitting* (Christian and Griffiths, 2016). If we were to have extensive, completely mistake-free data drawn from a perfectly representative sample, and a precise definition of exactly what needed to be evaluated, then the best approach would be using the most complex model available. In this ideal situation the only problem should be the presence of correlations among the different pieces of information used for building the predictive model. It is a well-known problem in statistics where it is defined as the *collinearity of the regressors* (Dormann et al., 2013). In the paradigmatic case of a dependent variable Y to be predicted by a set of independent variables Xs (regressors), the existence of mutual correlations between Xs creates a fundamental indeterminacy of the resulting model (Krishnan et al., 2007), causing unpredictable errors in subsequent applications. In any case, we can imagine that in an ideal case these problems can be faced by a preliminary factorization of the data set into mutually independent components (Xie and Kalivas, 1997), or by any other technique of prioritizing variable orders (Alhamzawi and Ali, 2018). Nevertheless, in a real situation that is far from being ideal, if we try to fit a given model to the actual data, a certain risk of incurring overfitting exists. In other words, overfitting poses a danger any time we are dealing with mismeasurement or environments that are vague, ambiguous, and ill-defined, as is commonly the case in a working setting and in everyday life due to the complex surroundings we live in. It is true that including more factors in a model will always, by definition, produce a better fit for our data. However, a better fit for the available data does not necessarily imply improved ability to generalize and thus better predict future cases that, by definition, are not available at the time the predictive model itself is constructed (Diaconis and Mazur, 2003; Christian and Griffiths, 2016).

Consequently, because any decision has to do with some kind of prediction of future outcomes, a better fit on actual data does not necessarily lead to a better decision. A model that is too simple can fail to capture essential aspects of the phenomenon studied; on the other hand, a model too complicated can become oversensitive to the particular adopted sample. Since the model is finely tuned to a specific data set, the resulting oversensitivity ends up intermingling a mixture of both general and idiosyncratic information relative to the specific sample, generating highly variable and consequently poor solutions. In this respect, it is worth noting that while the usual practice of splitting the whole data set into a "training" (from where the model is built) and a "test" set can be very wise, it only partially solves the overfitting problem. Successful models in science stem from the clear division of information into "sloppy" and "stiff" parts (in the jargon of data analysis). By focusing on solid information (usually consisting of very few control parameters) and leaving out the rest, it is possible to predict the behavior of very complex systems with good, and sometimes excellent, approximation (Transtrum et al., 2015; Giuliani, 2018; Ho et al., 2020). The inclusion of "sloppy" details marginally improves the adaptation of the model to the experimental data on which it is built, but at the expense of its generalization capacity.

It is possible to see the analogy with individual reasoning: an individual who excessively relies on an analytical approach, collecting as much data as possible and analyzing all details for their decision, will spend lots of energy in this effort and not necessarily select the best possible action. Being too analytical could lead to focusing specifically on details and losing the ability to consider the general purpose of the problem.

In other words, it is indeed true that knowledge originates from both deduction (move from the general to the specific), and induction (move from the specific, eventually many different specifics, to the general), but also strongly relies on *abduction* (Figueiredo, 2021). Abduction is the process of reasoning that goes from "the particular" to the general, but "the particular" is not a huge collection of data prepared to be analyzed by statistics, algorithms, and models. In abduction, "this particular" is a small set of specific and significant data that anchors our reasoning. It is the process used by doctors when they diagnose (Bird, 2010), by detectives when they try to resolve a criminal case, and by experts when they work. A paradigmatic example of the power of abductive reasoning in contrast to fully quantitative methods grounded in machine-learning is presented by Beaulieu-Jones, who shows the evidence of clinically driven decisions, based on heuristics approaches emerging from the personal expertise of the clinicians (Beaulieu-Jones et al., 2021).

From our point of view, the decision-maker who relies too much on the actual data can incur the dilemma of deciding based only on a specific sample of a much wider "population", losing generalization power and possibly leading to a worse prediction and thus worse decisions, or even to inaction (Diaconis and Mazur, 2003). We can imagine a kind of threshold beyond which the logical and analytical approach of collecting and analyzing data becomes counterproductive. This happens because we stop considering properties "common" to a certain class of problems and start to model the singularities of the particular reference set that have no equivalent outside the narrow scope from which the data derives (Transtrum et al., 2015).

Being aware of this phenomenon could change our approach to the decision in some way: mitigation of the previously described drawback could derive from the adoption of a behavior that not only relies on analytical and logical reasoning built on collected information, but also considers other subjective components. Although these aspects are difficult to express, extract or demonstrate with objective data, they could, in some way, be formalizable which is one of the novelties introduced by this work.

Foundational elements for a mathematical model of awareness

The development of a mathematical model of awareness can be addressed by starting from different perspectives and considerations (Friston, 2010). In this study we focused on philosophical, logical, cognitive and behavioral aspects; we aim to formally identify information flows and learning mechanisms that can allow individuals to initiate a process of increasing awareness instead of focusing on measuring conscious states and processes at the neural level (Modica and Rustichini, 1994; Heifetz et al., 2013; Karni and Vierø, 2017; Halpern and Piermont, 2020).

The first interesting contribution to the definition of a model of awareness can be found in philosophy and logic. According to Modica and Rustichini, a subject is *certain* of something when they know if it is true or false and *uncertain* when they know not its truth-value, and the subject's awareness of this not knowing is "conscious" uncertainty. On the other hand, a subject is *unaware* of something when they know not its truth value and is incognizant of their not knowing—they cannot perceive the object of knowledge, they are unable to mentally grasp it (Modica and Rustichini, 1999). The authors define *awareness* as the opposite of unawareness. Awareness includes both certainty and uncertainty, claiming that the concept of unawareness is a source of "ignorance" and distinct from uncertainty. They then propose the axiom of symmetry requirement for awareness: an individual can be aware of a proposition φ if and only if they are aware of its negation (*not*-φ). In other words, φ and $\neg\varphi$ (*not*-φ) are either perceived together or not at all. A logical definition of awareness is given in terms of knowledge: assuming $A\varphi$ means "the individual is aware of the propositional φ" and $K\varphi$ "the individual knows φ", the logical formulation of awareness is:

$$A\varphi = K\varphi \vee (\neg K\varphi \wedge K(\neg K\varphi)) \ \forall\varphi$$

which is PL-equivalent to:

$$A\varphi = K\varphi \lor K(\neg K\varphi)\ \forall\varphi$$

where PL stands for propositional logic. A subject who is aware of fewer things than another is not at all necessarily less capable of logical reasoning, those are two separate concepts. Nevertheless, it is possible the individual may not be capable of making some deductions precisely due to their unawareness: for example, if they are unaware of q they will not be able to deduce the concept of "*p implies q*" from knowledge of p, which would otherwise be doable to one who is aware of q. Therefore, awareness can improve people's decisions.

These aspects are relevant when choosing the right model class to characterize the processes that lead to increasing awareness—in our research that being the Markov Decision Process which shall be described in the model section. Notably, it assumes that increased awareness is the result of personal effort and clear focus with this specific objective in mind. Further, becoming aware reduces ignorance even if it does not reduce uncertainty, something intrinsically considered in the stochasticity of the model.

Awareness is a term that is often interchangeably used in different contexts, however, the literature dealing with awareness can be organized around three core concepts (Carden et al., 2022). The first is *cognitive awareness* (Papaleontiou-Louca, 2003) which identifies awareness as an accurate and deep understanding of an individual's perception, thinking, and actions. The second perspective argues that awareness exists on the multiple levels of consciousness and unconsciousness. Here, we consider awareness as an end-stage experience that results from the filtering and processing of several possible experiences happening simultaneously in our bodies and brains (Vaneechoutte, 2000). The third definition considers awareness with regard to recognizing others' feelings (Beck et al., 2004) and taking into account one's impact on others.

Further, whilst dealing with an individual's self-awareness, which is a relevant component of the developed model, two typologies are identified in the literature: "intra" and "inter" personal self-awareness (Carden et al., 2022). These are linked to an individual's own internal state and their impact on others, respectively, which can be further broken down into seven different components.

(1) *Beliefs and Values*. Beliefs Refer to the Conviction or Acceptance That Some Propositions About the World Are True, Especially Without any Proof, Whereas Values Denote the Hierarchical, Dynamic, and Abstract Attribution of a Degree of Importance to Something, Reflecting One's Judgment About What Is Important in Life. (2) *Internal Mental State* That, in Turn, Includes the sub-Components of *Feelings and Emotions* and *Thoughts and Cognitions* (Scherer, 2005; Wessinger and Clapham, 2009). (3) *Physical Sensations* Corresponding to Physiological Responses Manifested as Reactions in the Body. (4) *Personality Traits*, Reflecting the Individual's Stable and Consistent Characteristic Patterns of Thoughts, Feeling, and Behaviors. (5) *Motivations and Desires*, Which Are Related to Personal Drivers and Mental Directions. (6) *Behavior* Corresponds to an Action That Others see or Hear Individuals Displaying, Thus It Is an Interpersonal Component. (7) *Other Perception* Corresponds to how an Individual Is Perceived by Others, and the Ability to "Receive" Feedback. In Conclusion Self-Awareness can be Defined as:

> *Self-awareness consists of a range of components, which can be developed through focus, evaluation, and feedback, and provides an individual with an awareness of their internal state (emotions, cognitions, physiological responses) that drives their behaviors (beliefs, values, and motivations) and an awareness of how this impacts and influences others.*

In developing a mathematical model, these findings allow self-awareness as a fundamental factor of awareness to be built in, including emotions, cognitive processes, motivations, believes, evaluations and feedback. All these elements can easily be embedded into a Markov Decision Process.

All the factors outlined above have a heavy impact on many other aspects of our social life. The ability to use tacit knowledge and intuition is common and necessary anytime people need to make decisions in complex environments that are future-oriented, highly uncertain, difficult to forecast and lacking in information (Mintzberg, 2000). For example, in the field of leadership and management, rational decision-making skills are required to enable processing available information clearly and logically and thus permitting accurate perception and interpretation of the incoming events, which can sometimes lead to creative and innovative solutions (Prietula and Simon, 1989). Nevertheless, apart from this type of knowledge is essential to consider that managers routinely make decisions based on knowledge grounded in experience and could use intuitive decision strategies, especially under high-stress conditions (Sayegh et al., 2004). Tacit knowledge—the work-related practical know-how acquired through direct experience and instrumental in achieving goals important to the holder (Brockmann and Anthony, 2002)—is not easily recognized or acknowledged, but it can be a key factor in enhancing the quality of strategic decisions.

The sensible application of tacit knowledge can partially fill information gaps ameliorating the efficiency of the decision process (Brockmann and Anthony, 1998). Further, self-awareness is now seen as a critical component in leadership and career success, joining the skills of team interaction development, effective coordination and collaboration (Dierdorff and Rubin, 2015), and non-conflictual and sensible leadership (Axelrod, 2005).

Developing a mathematical model of awareness

In this section we introduce a mathematical model that can incorporate the main factors summarized in the previous sections. According to the statements discussed in Section Decision-making processes: an overview, we have applied the class of models referred to as Markov Decision Processes, which have been deemed suitable to describe human decision-making under uncertainty (Rangel et al., 2008), including autopiloted decisions (Landry et al., 2021) and addictive behavior (Mocenni et al., 2019).

Sequential decision models and Markov Decision Processes

Each day people make many decisions that have both immediate and long-term consequences. Decisions must not be made in isolation, today's decisions impact tomorrow's and tomorrow's the day after; if one does not account for the relationship between present and future decisions and present and future outcomes, one may not achieve overall good performance. The *Sequential Decision Models* (SDMs) consider both outcomes of current decisions and future decision-making opportunities under some kind of uncertainty (Puterman, 2014).

In a sequential decision-making model at a specified point in time (that is also called "decision epoch") the *Decision-maker* (DM) observes the current state of the system, and based on this state, chooses an action among the ones available in that state. The choice produces two results: the DM receives an immediate reward (or incurs a cost) and the system evolves to a possibly new state in the next decision epoch. At this subsequent point in time, the DM faces a similar problem as schematized in Figure 2. The key ingredients of this sequential decision model are:

- A set of decision epochs;
- A set of states;
- A set of available actions (which can be different in different states);
- A reward (or cost) function depending on state and action;
- A set of state transition probabilities depending on state and action.

We usually assume that the DM knows all these elements at the time of each decision, thus they constitute the amount of explicit information available.

At each decision epoch, the DM performs a choice, and they can have a *policy* that provides the most favorable action in each possible state; the implementation of a policy generates a sequence of rewards (or costs). The sequential decision problem consists in finding a policy before the first decision epoch, which maximizes a function of the rewards' sequence (or minimizes the costs' sequence). A policy that accomplishes this task is defined as optimal and relative to the specific individual and function considered.

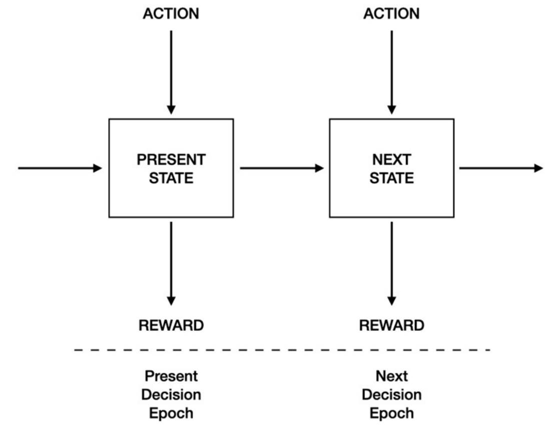


FIGURE 2
Schematic representation of a Sequential Decision Model.

Markov Decision Processes (MDPs) are a particular class of SDMs in which actions, rewards, and transition probabilities solely depend on the current state, and not on occupied states and actions chosen in the past. In other words, the current state incorporates the entirety of the DM's past: it is the result of all their previous decisions, related outcomes, and experience gained from them. In this work, we try to characterize some dynamics of the awareness-raising process. Therefore, we assume awareness is a dynamic process (characterized by a sequence of states with a certain dynamic in time) involving the DM's experiences, filtered by his perspective, beliefs, values, actual state, and choices, through different reasoning attitudes. Moreover, in an MDP, we simultaneously have the presence of a decision-maker's choice and uncertainty regarding its outcomes, as always happens with our decisions due to uncontrollable factors.

This model should be a good trade-off between realism and simplicity: broad enough to account for realistic sequential decision-making problems while simple enough to allow it to be understood and applied by different kinds of practitioners.

It is worth stressing that MDP, even if incorporates in a given state s_t the DM's entire past, represents a future evolution uniquely dependent on the current state and is completely independent of the particular trajectory that reaches state s_t at time t. Only *a posteriori* reflection by a DM on the trajectory of past experience builds (in the long run) both awareness and tacit knowledge. Here we recognize the presence of two dynamics with very different time scales. The short timescale of MDP, ending up in the terminal decision, and the long timescale (we can think of the physical analogy of a capacitor) that generates both awareness and tacit knowledge by reflecting on past stories of successes and failures that can, in turn, be of use for future decision processes.

Model formulation

Time and state

As mentioned above, the adopted model belongs to the framework of Markov Decision Processes (MDPs) and considers a discrete and finite time horizon of length T in which, to each *time-epoch*, t, corresponds to a moment of making a relevant decision—which needs some kind of reasoning process and is not purely automatic and routine. Since the life of an individual is limited, it is reasonable to consider a finite time horizon. The *state*, $s_t \in (0,1)$, of the individual is a representation of their level of awareness at each time-epoch t and belongs to the set S represented by the discrete closed interval (0,1) with a step size of 0.01. In this way, a unique definition of awareness is not explicitly given, which would be a very difficult task, it is simply said that awareness is a state of the individual which determines their well-being from a global point of view and has a considerable impact on their choices. This paper is mainly focused on discussing how awareness can be accounted for by a mathematical model more than giving an explicit definition of it. It introduces an attempt to model some mechanisms underlying the process of aware decision-making rather than quantifying the effective awareness of individuals in some way, which could be a herculean task. By considering an MDP, the current state incorporates the DM's complete history so that their awareness is a state, in some way, embodying all of the individual's past: from their personality, values, beliefs developed over their lifetime, to their education and past experiences.

Reasoning propensity

The *reasoning propensity*, $p_r \in (0,1)$ embeds the specific attitude in processing the information about the problem, and represents the trade-off between the two reasoning modalities: *analytical* and *intuitive*. This combination depends upon different individual factors like age, character, beliefs, values, desires, education, experience, and so on; it varies from individual to individual but can also change in the same individual throughout their lifetime. The reasoning propensity takes values in a continuum between the two extreme attitudes (Allinson and Hayes, 1996), called *intuitive* ($p_r = 0$) and *analytical* ($p_r = 1$), assuming in this way that both are always involved, to different degrees, in any decision. These two modalities refer to the dichotomy between rationality and intuition, as Section Decision-making processes: an overview brings to light.

Policy and decision

The reasoning propensity affects the *policy*, μ of the individual. Generally speaking, a policy is a function that prescribes the action to make for each possible state at any time instant, and is represented by a matrix of dimensions [|S| x T]. It can be somewhat complicated to shape different situations. Therefore, the policy turns out in the *decision*, u_t, which belongs to the open interval $U = (0,1)$, so that the more analytical the choice, the higher the value of u_t. We have that:

$$u_t = \mu\,(s_t, t)\ \forall\, s_t \in S \textit{ and } t = 1, \ldots, T$$

The choice leads to two results: the DM receives a reward, and the system possibly evolves to a new state.

State evolution and transition probability functions

The DM's state, s_t, evolves according to:

$$s_{t+1} = f(s_t, u_t, w_t)$$

That is, the future level of awareness of the individual depends on the current state, the choices they mak e, and its outcome, which is subject to some uncertainty represented by w_t, the *stochastic variable* related to a state transition. We assume, for simplicity, that the state can remain the same or increase and decrease by only one step, in this way w_t belongs to the set $W = \{1, 0, -1\}$, indicating, respectively, the possibility that the state increases, remains constant or decreases by making a decision u_t. The presence of uncertainty affecting the outcomes of the decision due to uncontrollable elements in the environment makes the state evolution and the rewards sequence stochastic. There exists a known *transition probability*, function of u_t, specified in a matrix P of dimensions [|U| x 3]. In particular, P has the form:

$$P = [P_1(u_t)\, P_0(u_t)\, P_{-1}\,(u_t)]$$

Each one of the three columns of P specifies the probability that the state increases, remains constant, and decreases, respectively. In other words:

- $w_t = 1$ *with probability* $P_1\,(u_t) \rightarrow$ *Forward probability*
- $w_t = 0$ *with probability* $P_0\,(u_t) \rightarrow$ *Stationary probability*
- $w_t = -1$ *with probability* $P_{-1}\,(u_t) \rightarrow$ *Backward probability*

In this way, the system dynamics can be re-written as:

$$s_{t+1} = s_t + w_t$$

Notice that all the elements in the matrix P are values representing a probability, and are subject to two conditions:

$$0 \leq P_w(u_t) \leq 1\ \forall\, w \in W \textit{ and } P_1(u_t) + P_0(u_t) + P_{-1}(u_t) = 1\ \forall\, u_t \in U$$

The *stationary probability* P_0 has been defined as a constant value, notably all simulations consider:

$$P_0(u_t) = 0.1\ \forall\, u_t \in U.$$

It incorporates the DM's resistance to change their level of awareness, and depending on their characteristics, can be considered "inertia". The *forward probability* results from the linear combination of two fixed functions exploiting the cases of intuitive and analytical reasoning.

Figure 3 shows the functions considered for the forward probability in the—only theoretical—cases of a complete analytical (Figure 3A) and a complete intuitive (Figure 3B) individual. The first one starts with a low value and then increases as the decision becomes more analytical. It reaches the maximum when the reasoning is highly analytical, and then, for bigger values of u_t, the probability starts to decrease. This is a representation of the *overfitting* phenomenon described in Section Decision-making processes: an overview, exploiting the fact that an excessively analytical approach to reasoning could also turn out to be counterproductive. The second function has an opposite behavior: the more intuitive the reasoning (the smaller u_t), the bigger the probability of increasing the state. This is because the individual thinks to have access to personal abilities, not related to cognition, allowing their level of awareness to increase by using an appropriate degree of intuition; this has to do with the personal confidence in tacit knowledge. It is possible to consider that a minimum level of analyticity is indispensable to understand the framework of the decision in this case; otherwise, intuition loses contact with the reality of the decisional problem, becoming only a fantasy. An excessively intuitive individual may act without considering the context from which the decisions come, and the decision could be ineffective. We must note that all these transition functions reflect the DM's different points of view; we are putting ourselves in the shoes of the individual. It is very difficult to consider a transition probability function that generically specifies the probability of increasing the state of individual awareness without depending on any such kind of assumption.

The two basic functions (Figures 3A,B) have been designed so as to represent the different theoretical assumptions exposed in Section Decision-making processes: an overview, including the drawbacks of being excessively analytical or intuitive. Certainly, different functions can be considered as long as they are capable to incorporate the same phenomena.

As we mentioned above, any real DM mixes, to some extent, these two modalities according to a personal proportion represented by their reasoning propensity p_r. The effective forward transition probability of the DM, i.e., the probability of increasing the state's level, is computed as the convex combination of the two fixed functions—forward transition probabilities in the only theoretical case of complete analyticity or intuitiveness of the individual—using the reasoning propensity p_r as coefficient:

$$P_1(u_t) = p_r P_1^{analytical}(u_t) + (1 - p_r) P_1^{intuitive}(u_t)$$

Figure 3C shows the influence of different values of p_r on the transition probability P, as described in the legend.

Finally, the *backward probability* is defined starting from the first two as:

$$P_{-1}(u_t) = 1 - (P_1(u_t) + P_0(u_t))$$

Rewards

The problem now arises of how to define a function that grants the individual a reward by selecting choice (u_t) instead of another and being in a certain state s_t. Is it important to maintain the focus on what are we trying to explain with the model; that is: investigate the dynamic underlying the process of awareness-raising. In fact, as exposed in Section Decision-making processes: an overview, the dynamic of awareness-raising emerges from personal effort and motivation. Moreover, as human agents, we are accustomed to operating with rewards that are so sparse we only experience them once or twice in a lifetime, if at all. Most goals of modern life—a good job, a house, a family, a happy life—are so abstract, complex, and far into the future that they do not provide useful reinforcement signals. Despite this, people continuously make choices in their lives, applying what psychologists call *intrinsic motivation* or *curiosity*. Motivation/curiosity have been used to explain the need to explore the environment and discover novel states. Similar to what also happens in reinforcement learning, intrinsic motivation/rewards become critical whenever extrinsic rewards are sparse (Pathak et al., 2017). In our case, intrinsic motivation refers to reaching higher states of individual awareness, which can be linked to reaching sparse, temporary, distant and extrinsic life goals.

Mathematically the reward function consists of two parts: a *stage reward* explicitly depending on state and choice, and implicitly on the stochastic variable $\mathrm{w_t}$. The second is a fixed *terminal reward*, $r(s_T)$, which the DM incurs at the last time-epoch T. The stage reward linearly depends on the current state and the choice, with constant and positive coefficients α and β:

$$r(s_t, u_t) = \alpha s_t - \beta u_t$$

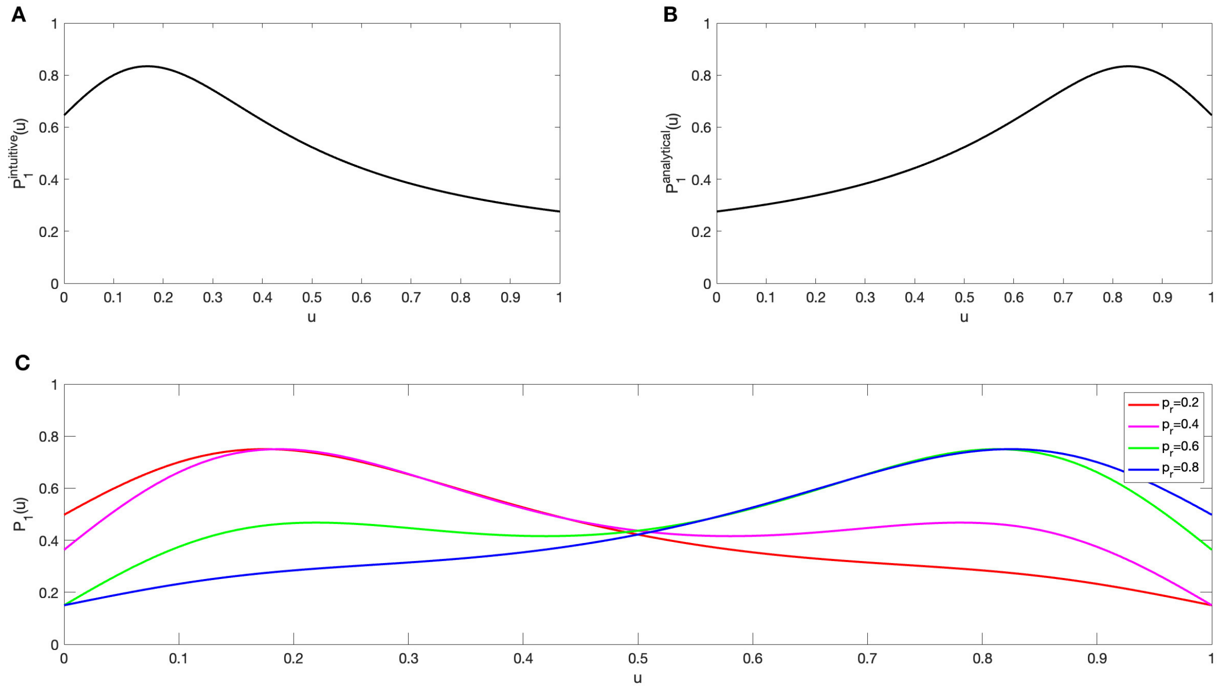


FIGURE 3
Forward transition probability. **(A, B)** indicate the forward transition probabilities of an intuitive and analytical individual, respectively. These two functions are linearly combined using the specific individual's reasoning propensity p_r. Some examples of forward transition probability functions for different p_r are shown in **(C)**.

It is reasonable to assume that the individual's current level of awareness has a positive influence on an individual's whole life, so living with a higher level of awareness can improve well-being on all levels (physical, psychological, emotional, and so on). Consequently, the equation incorporates a positive dependence on the current level of awareness so that the higher it is, the better the individual's life is overall. On the other hand, rational/analytical reasoning is resource-consuming because it requires the acquisition of some kind of data about the problem and the possible alternative solutions, and requires time to analyze and elaborate all the data. Intuitive reasoning, as also revealed in Pascal's thought, is *effortless* and not resource-consuming. Therefore, the more the decision implies analytical reasoning the more resources it needs in terms of time, personal energy, and monetary resources. This translates into a negative dependence of the reward on the choice u_t, because the higher u_t is the more analytical the reasoning of the DM, and so the more resources consumed. Although the current version of the model assumes a linear form for the step-reward function, to explain the thoughts behind its formulation in a simple way other typologies functions are possible and may be more suitable.

For the same reasons set out in the previous point, we can assume that the terminal reward the DM incurs at time T increases with the value of the ending state (Figure 4). It has been considered an exponential function that "tries to push" the final state as high as possible, providing a considerably different reward between ending in a "high" state rather than in a "low" state.

Here we can also notice how tacit knowledge can be thought of as deriving from the sedimentation of past cases into an "experience capacitor". The terminal rewards derived from past cases lose their specific reference to the actual situation from which they stemmed, and contribute to the creation of a "good practices" repository no longer linked to a particular situation but a "broad spectrum of cases".

Future weights

Equal rewards at different time-instants have a different value for the individual. Therefore, factor δ weigh ing future rewards has been introduced. Different applications and aspects referred to this consideration are studied in detail in Section A model extension: Including individual emotions.

Backward penalty

In the end, we considered a vector γ of dimensions [3 x 1] to give possible different weights to the cases of increasing, maintaining constant, or decreasing the state, respectively. We assume different values of γ in different simulations,

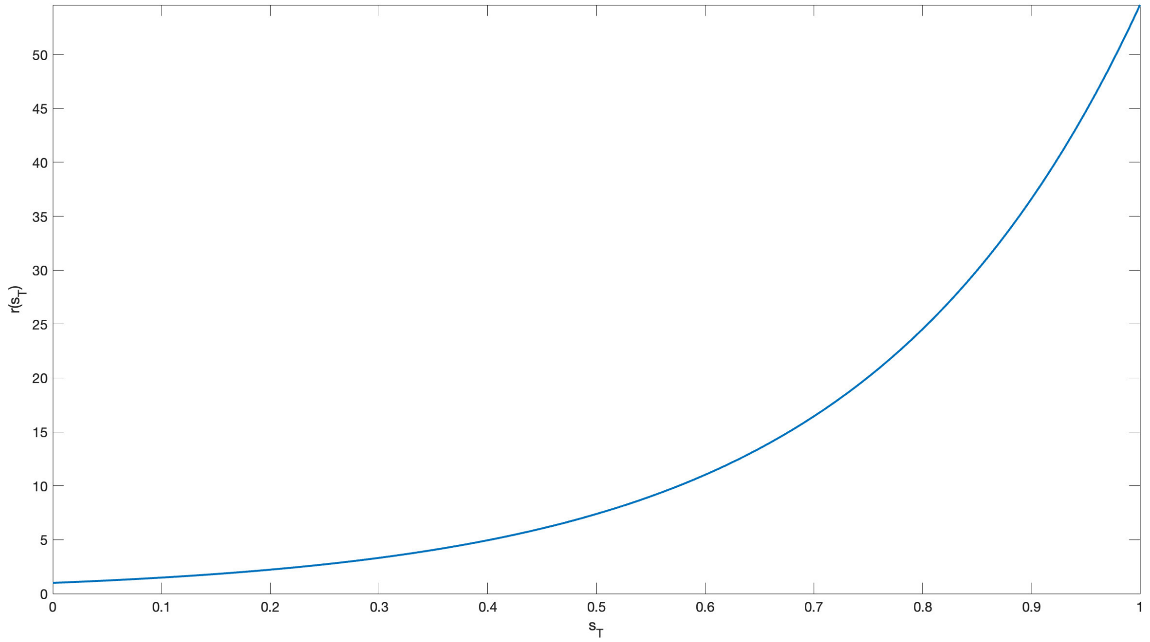


FIGURE 4
Terminal reward as a function of the level of awareness at the final time instant.

accounting for different attitudes of the DM—for example strongly penalizing the eventuality of decreasing awareness.

Habitual decisions

Once the general structure of the model and the meaning of its parameters are defined, it is possible to consider how the resulting model can be applied to different situations shaped through different policies. In the following sections we present two ways to find suitable policies aimed at solving the decision problem. The first, proposed in this paragraph, refers to the most basic and simplest mechanism governing an individual's *habitual decisions* (or *choices*), the ones made with little to no effort and without conscious control (Landry et al., 2021). They consequently assume that the DM does not have any self-awareness, so that decisions spring only from their habits as automatic, non-conscious mechanisms.

As mentioned in the previous section, the DM has their own reasoning propensity $\mathrm{p_r}$ indicating how much the decision-making process is intuitive or analytical. It is possible to assume that this is the only characteristic governing habitual decisions, considering a naïve policy, that is accordingly called *habitual* or *usual policy*, defined as:

$$\mu\left(s_t, t\right) \ = \ e_t \ \forall s_t \in S \ and \ t \ = \ 1, \ldots, T$$

Where e_t is normally distributed with mean p_r and standard deviation σ fixed to a constant value, for example 0.3, represents the fact that any individual's decision always encompasses the processing of information regarding a problem in a similar way, more or less analytically. However, a certain variation in the choices has been considered around the value representing the propensity of reasoning, supplied by other uncontrollable contextual factors which are the real drivers of the decision, making the individual unaware of being able to effectively decide the value to choose. These factors represent a source of uncertainty, influencing the choice, and can drive it far from the effective p_r, highlighting the case in which people are not synchronized with their effective reasoning propensity.

Ultimately, this policy's structure shapes the case of the DM's unaware decisions. It is possible to see some conceptual similarities to the UMDPs (Unaware Markov Decision Processes) which represent an attempt to introduce the concept of unawareness in the framework of Markov processes (Halpern et al., 2010). A common idea is the restriction of the set of possible actions, even if implemented in different ways, reflecting the unawareness of the DM regarding an entire set of possible actions. In the policy described above, beyond this kind of unawareness, the individual is also unaware of their effective reasoning propensity, assuming that the effective choice randomly selects a value around it. In this way, the unaware choices are not completely random but reflect a kind of coherence of the individual and, on the other hand, shape an unawareness of what is the real p_r.

In this work, this structure is mainly applied as a term of comparison to evaluate the effect of introducing an individual's self-awareness on the choices.

Self-aware decisions

The second structure that is proposed represents a first attempt to incorporate the concept of self-awareness in the process. If we think about self-awareness, we could imagine that it is an element deriving from some kind of self-observation—a "third person" perspective from a metacognitive point of view (Drigas and Mitsea, 2020)—and that has a consequent impact on the action/decision. Mathematically it can be represented by a *feedback loop*, according to the logical representation of Figure 5.

Accordingly, a component that modifies the policy has been introduced by additionally observing the form of the transition probability function, current state, and time epoch. In this way, the DM is allowed to modify their usual, automatic process by shifting from their habitual to a new policy that mathematically results from a maximization process. This introduces the possibility of mitigating the habitual tendencies of the individual by modifying the policy.

This feedback is mathematically embedded in an optimization process, intended to maximize the sequence of rewards. Due to the linear dependence of the reward on the level of awareness, it is also thus modeling the fact that self-awareness results from a personal effort.

Computation of the optimal policy

As previously mentioned, a policy is *optimal* when it maximizes a certain objective function which, in this case, is the cumulative sum of the rewards incurred at each time epoch. One of the methods that can be applied to compute the optimal policy in an MDP is the *Stochastic Dynamic Programming* (SDP) algorithm, choosing the action which maximizes the sum of the current reward and the expected future rewards at each stage. Mathematically we can say that the following problem must be solved:

$$\begin{aligned} \max_{\mu} E & \left[\sum_{t=0}^{T-1} r\left(s_t, \mu\left(s_t, t\right), w_t\right) + r(s_T) \right] \\ s.t.\ s_{t+1} & = f\left(s_t, \mu\left(s_t, t\right), w_t\right)\ t = 1, \ldots, T \end{aligned}$$

Considering a finite time horizon of length T, a decision is *not* made at time T, so that the DM's last choice is at time T-1, and the final time instant is used to fix a terminal reward the DM incurs at time T, $r(s_T)$. From there it is possible to recursively reconstruct the optimal policy by exploiting Bellman's *Principle of Optimality* (Bellman and Drayfus, 2015) which affirms that "*an optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting for the first decision".* The original problem can be decomposed into a recursive series of easier sub-problems, considering a shorter time horizon from τ to T and a given initial state :

$$\begin{aligned} V_{\tau}\left(\underline{s}\right) & = \max_{\mu_{\tau} \ldots \mu_{T-1}} \left[\sum_{t=\tau}^{T-1} r\left(s_t, \mu\left(s_t, t\right), w_t\right) + r(s_T) \right] \\ s.t.\ s_{t+1} & = f\left(s_t, \mu\left(s_t, t\right), w_t\right)\ s_{\tau} = \underline{s}\,,\ t = \tau, \ldots, T \end{aligned}$$

where $V_{\tau}\left(\underline{s}\right)$ is the value function that indicates the optimal reward cumulatively obtained considering the sub-problem with time horizon τ, …, T and initial state. Starting with $\tau = T$-1 and then decreasing the value of one unit each time, it is possible to recursively calculate the optimal policy for which the current optimal value can be seen as the sum of the expected stage reward and the expected optimal value function at the next time instant:

$$V_{\tau}\left(\underline{s}\right) = r\left(\underline{s},\ \mu\left(\underline{s}, \tau\right)\right) + V_{\tau+1}\left(f\left(\underline{s}, \mu\left(\underline{s}, \tau\right), w_t\right)\right)$$

This expression in our case can be expanded considering that w_t can assume three values with probabilities specified in matrix P. So, it becomes:

$$\begin{aligned} V_{\tau}\left(\underline{s}\right) = & \ r\left(\underline{s},\ \mu\left(\underline{s}, \tau\right)\right) + \delta\left[\gamma_1 V\left(\underline{s} + 1\right) P\left(\mu\left(\underline{s}, \tau\right), 1\right)\right. \\ & \left. + \gamma_2 V\left(\underline{s}\right) P\left(\mu\left(\underline{s}, \tau\right), 2\right) + \gamma_3 V\left(\underline{s} - 1\right) P\left(\mu\left(\underline{s}, \tau\right), 3\right)\right], \end{aligned}$$

where $0 < \delta < 1$ is the weight given to the next-instant reward, and $\gamma = [\gamma_1, \gamma_2, \gamma_3]$ is the vector of the different weights given to the possibility of increasing, remaining constant, and decreasing the next state, respectively. Each constant coefficient γ_i belongs to (0,1).

It is worth remembering that the transition probabilities explained in matrix P depend on the reasoning propensity p_r of the individual, and this determines the effective shape of the curve representing forward, stationary, and backward probabilities as a function of the decision (u_t) suggested by the policy.

A model extension: Including individual emotions

Immediate emotions (also called *visceral factors* in economics) play a critical role in the intertemporal choice modifying the utility of an action, leading people to behave in ways that appear to greatly discount the future, ways that individuals themselves can sometimes see as contrary to their own self-interest (Loewenstein, 2000).

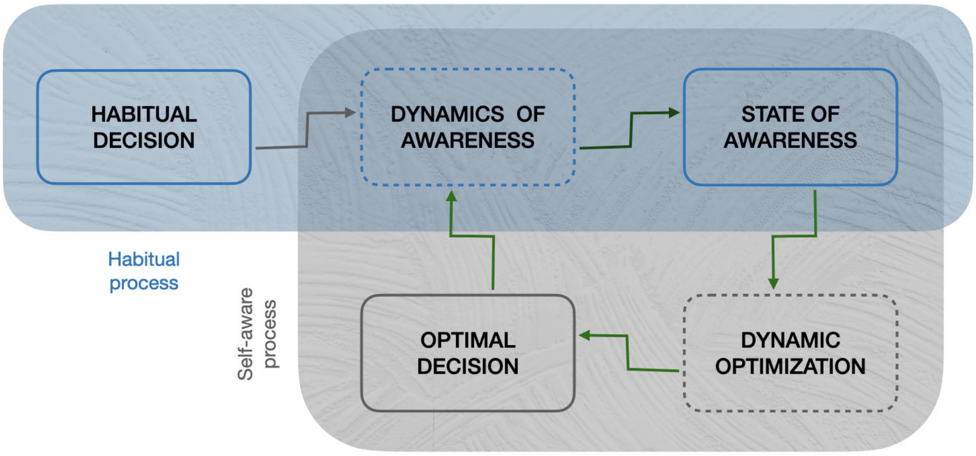


FIGURE 5
A schematic logical representation of the model. The blue (green) circle indicates the structure in the case of *habitual* (*self-aware* or equivalently *optimal*) decisions and their intersection. The dashed borders of the two blocks relative to the dynamics symbolize the uncertain factors impacting the dynamic in time.

In this description, we can identify three basic aspects that could help model emotions: their relationship with time, with perception (utility) or a reward, and the fact that they could also be counterproductive. In the proposed model the ideal place to insert emotions seems to be in factor δ that weights future rewards. It permits connecting immediate emotions to the perception of the future and the value of the rewards.

We especially claim that emotions do not necessarily hurt an individual's choice but can also "help" them. We can equate emotions to the role played by "temperature" in simulated annealing optimization models (Bertsimas and Tsitsiklis, 1993); in order to escape eventual local minima during the optimization process the simulated annealing algorithms allow for a certain degree of stochasticity that could inhibit the system to take the "most convenient move" during the optimization process. This mirrors the role of temperature that can make the system exit from a potential hole, the temperature (i.e. the degree of stochasticity) decreases during the process and goes to a minimum near the end of the process so as to not destroy the reach of the optimal solution. It is worth noting that, at odds with simulated annealing, our model does not incorporate an explicit decreasing trend of temperature (contribution of emotion) but an equivalent effect is reached by introducing a dependence on awareness's dynamic along the process that in turn can make the emotions somewhat less relevant.

The entity of the role played by emotions depends on the level of awareness of the individual. At a low level of awareness, emotions prevail on individual reasoning, so one is completely driven by choices in search of instant gratification (independent of the reach of the actual target). In this condition, the future will have very little weight on one's choice, which can be detrimental because people behave in a way that is contrary to their self-interest. Contrarily, this dynamic is not present when the individual reaches a high level of awareness in which one could consider emotions freely and peacefully, and could, in some way enhance a benefit from the choice.

Another aspect to considered in the computation of δ is the relationship between future weight of choice and the age of the individual, in such a way that the older the individual, the bigger the weight they give to future rewards. An older individual with less time to live consequently gives more importance to each possible moment in the future; in contrast, a younger individual could weigh the present with less consideration of the future.

Mathematically, this additional extension changes the equation defined in the paragraph regarding the computation of the optimal policy, which becomes:

$$\begin{aligned} V_\tau\left(\underline{\mathrm{s}}\right) = {} & r\left(\underline{\mathrm{s}},\ \mu\left(,\tau\right)\right) + \delta\left(\underline{\mathrm{s}}\right)\frac{t}{T}\left[\gamma_1 V\left(\underline{\mathrm{s}}+1\right)P(\mu\left(\underline{\mathrm{s}},\tau\right),1\right) \\ & +\gamma_2 V\left(\underline{\mathrm{s}}\right)P(\mu\left(\underline{\mathrm{s}},\tau\right),2\right) + \gamma_3 V\left(\underline{\mathrm{s}}-1\right)P(\mu\left(\underline{\mathrm{s}},\tau\right),3\right)\right], \end{aligned}$$

with time horizon $\tau, \ldots, T$ and an initial (known) state $\underline{\mathrm{s}}$.

The term $\delta\,(s)$ is introduced in the model to insert an emotional component. It indicates a modification of the structure of δ which, until now, was a constant value but is now considered as a function of the state (see Figure 7A). Moreover, it reports a linear dependence on time t, where the term $\frac{1}{T}$ is a scale factor. Summarizing, the new term embeds the impact of emotions in the decisions as a factor which enforces the expected value of the future reward when either the awareness level increases, or the time horizon reaches its maximum, or both.

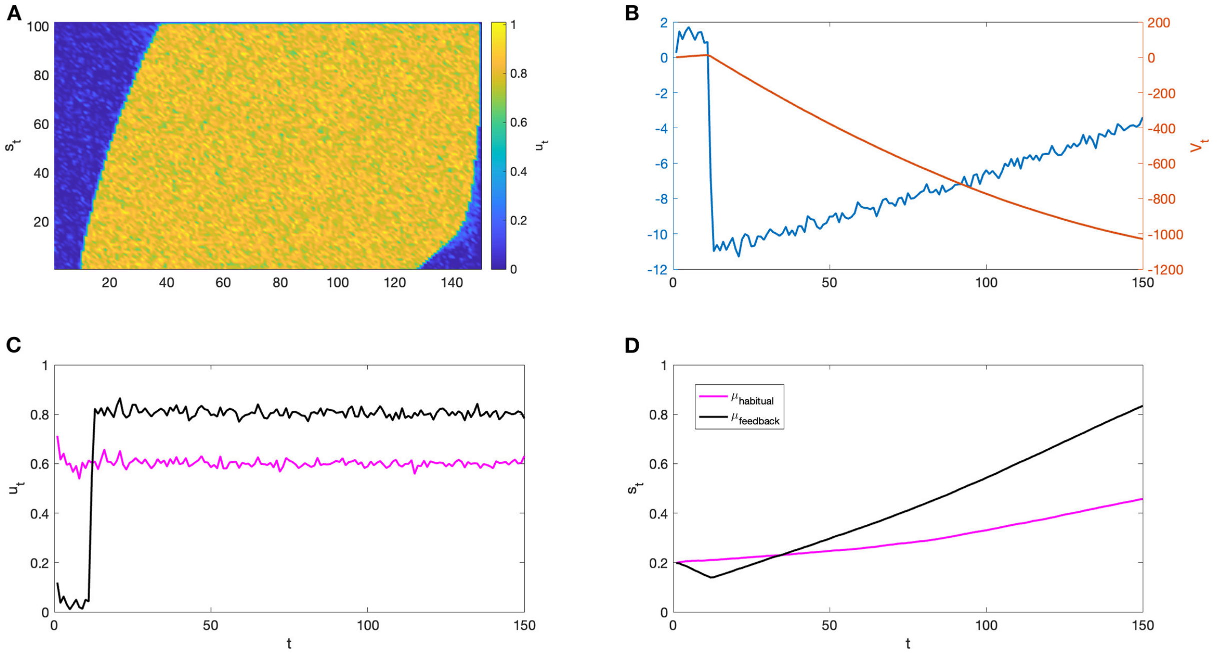


FIGURE 6
Some results of the simulations. We considered an individual with $p_r = 0.6$, low noise on the policy ($\sigma = 0.08$), and a low initial state ($s_0 = 0.2$). The other model's parameters were fixed to $\alpha = 1$, $\beta = 1.5$, $\gamma = [3\ 1\ 0.1]$, and $\delta = 0.75$. **(A)** reports the matrix of the optimal policy computed in the presence of the feedback loop: it indicates the decision u_t (indicated by the color) to perform for each combination of time epoch and state. **(B)** reports the step (blue) and cumulative (red) rewards in the presence of feedback. **(C, D)** highlight the different evolution of the states and decisions considering habitual (magenta) and self-aware (black) behavior, respectively.

Numerical results

The next step was to carry out numerical simulations to apply the different structures corresponding to the habitual and self-aware policies outlined in the previous section in order to evaluate the evolution of the dynamic. To do this, it must be also specified:

- The *number of simulations*, N, to perform. Each of them with a time horizon of length T.
- An *initial state* s_0 for each simulation. It can be fixed to a particular value to evaluate the dynamic starting with a specific level of the state or can be computed as a random value extracted from a discrete, uniform distribution that takes values from 0 to 1.
- Notice that t_0 and s_0 are related to the instant when the observation starts, they are not intended as absolute values yet always have a relative connotation.
- For each time instant in each simulation we need a *realization* of the random transition variable w_t which takes values in $W = \{1, 0, -1\}$ according to the probability functions in P evaluated at u_t; in fact, the DM implements, at each time, a choice according to the policy μ they choose ($u_t = \mu(s_t,t)$).

Performing $N = 3{,}000$ simulations and analyzing the average value in all of them, we can see what happens to the trends of state and rewards.

Some numerical results are reported in Figure 6, obtained by considering an individual with $p_r = 0.6$. The habitual policy for such a kind of individual is constantly centered around 0.6 with, in this case, low noise. It means that the choice of the individual always has roughly 60 percent analyticity of reasoning. From the optimal policy's matrix (Figure 6A), one can see that the optimal policy suggests starting with a low level of u_t and then increase it to 0.8 (Figure 6C). The lack of observation of transition probabilities and the level of state creates, in the habitual case, a decreased rise in state, which at ending time reaches 0.5, whereas with the introduction of feedback the state is able to saturate near the maximum state (Figure 6D). We have chosen this particular case to discuss in light of the phenomenon that the state initially decreases in the presence of feedback. It is generally possible to notice that the state with the feedback loop monotonously increases and is higher with respect to the other. These results are explained in our first publication on that topic (Bizzarri and Mocenni, 2022), where the embryonic idea of comparing habitual and optimal strategies in human decision-making has been presented, while the mathematical model, including the concept of overfitting, tacit knowledge and emotion, have been

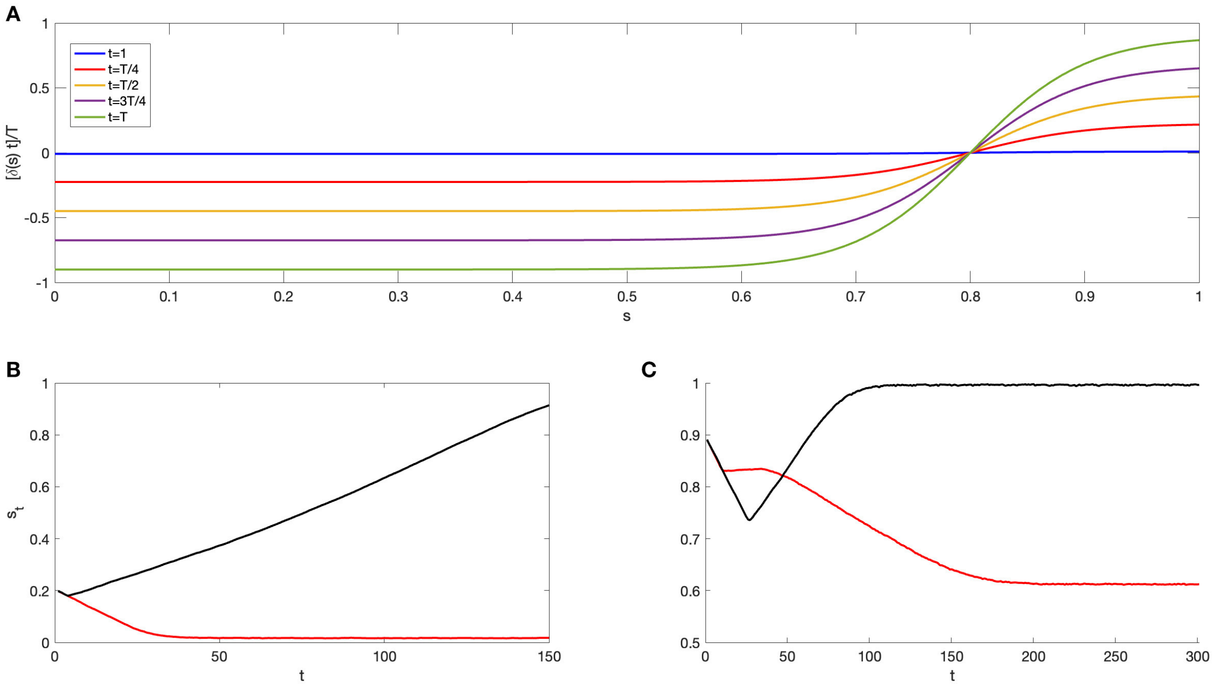


FIGURE 7
Analytical individual with emotions. **(A)** reports the new term $\delta(s)\frac{t}{T}$ defined as a function of the state s_t for different specific values of time: $t = \{t_0, T/4, T/2, 3T/4, T\}$. **(B, C)** analyze the effect of emotions on an individual with $p_r = 0.8$, starting with a low ($s_0 = 0.2$) and a high ($s_0 = 0.9$) initial state, respectively. The red line indicates the behavior in presence of emotion whereas the black line in absence of emotion (considering a constant $\delta = 0.75$). In both cases, the feedback loop is included, modeling self-aware decisions. The other model parameters are fixed to $\alpha = 10$, $\beta = 20$, $\gamma = (3.7, 1, 0.01)$, $\sigma = 0.08$.

introduced in the present paper for the first time. However, by considering a different parameter setup we can notice that even if the state temporarily decreases, the presence of a feedback loop allows for a change in the trend, reaching values that are even higher than in the case of habitual behavior. The step reward has a similar trend with an initial decrease and then a more rapid increase than in the habitual case (Figure 6B—blue line).

Including emotions in the model

It is possible to evaluate the impact of embedding emotions of the individual by performing some simulations and correspondingly modifying the computation of the optimal policy in the presence of a feedback loop, as exposed in Section 4.5.

It is possible to observe that a highly analytical individual starting from a low state manifests a decrease of the state in the presence of an emotional factor, as described in Figure 7B. This is due to the new form of $\delta(s_t)$ (Figure 7A), which has a negative value for a low state of awareness, claiming that in this case the presence of emotions greatly effects future discounting which could also turn out as harmful (in case of negative values of δ). On the contrary, after gaining a certain level of awareness, δ starts to increase reaching a value near 1, meaning that an individual with a high level of awareness does not make any distinction between the present and future.

Figure 7C demonstrates the possibility that emotions could enhance the state evolution: in this case, emotions are helpful in increasing the state evolution over transient times, until it stabilizes at a constant value ($s_t \sim 0.7$). This behavior appears when considering a longer time horizon, where T is set at 300.

The helpful effect of emotions starting from a high s_0 is more evident in the following when considering an intuitive individual (Figure 8C).

In the case of a highly intuitive individual, the additional emotive aspect creates an oscillatory behavior when considered at a lower initial state (Figures 8A,B). This oscillation makes it impossible to choose a stable value of u_t, which oscillates between high and low values. Consequently, it stops the growth of the state.

As also highlighted in the analytical case, the presence of emotions at high states enhances the evolution of the state that increases faster than without emotions (see Panel C, where the red line is over the black one).

Discussion

The aim of this research was to investigate how to integrate facets like tacit knowledge, intuition, emotions, awareness,

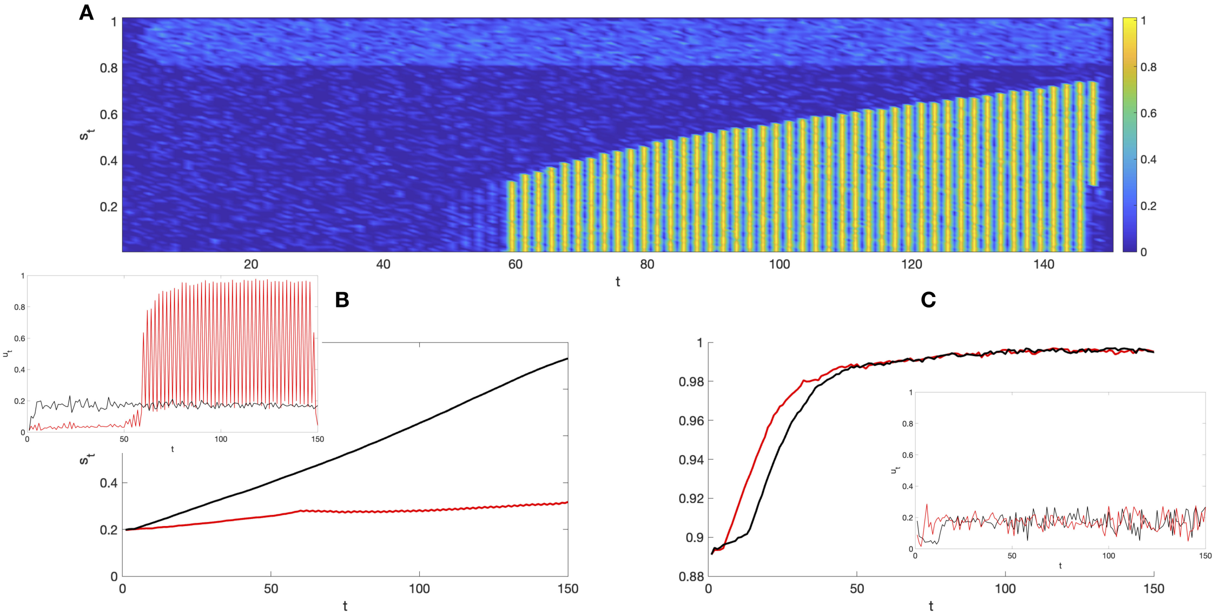


FIGURE 8
Intuitive individual with emotions. Now we shall consider an individual with $p_r = 0.2$, analyzing, as before, the trends due to emotion starting from a low initial state. **(A, B)** analyze the effect of emotions starting low ($s_0 = 0.2$). **(A)** reports the optimal policy whereas **(B)** demonstrates the state evolution and, in the insert, the effective choice over time. **(C)** considers a high ($s_0 = 0.9$) initial state and shows the state dynamic and, in the insert, the efficient choice. The red line indicates behavior in the presence of emotion whereas the black line in absence of emotion (considering a constant $\delta = 0.75$). In both cases the feedback loop is included, modeling self-aware decisions. The other model parameters are the same used in Figure 7.

and self-awareness into a mathematical model of decision-making, going beyond the classical analytical perspective. These factors have been considered within the framework of a richer conception of the decision-maker, and their multi-faceted effects are intensely analyzed. Even though all of them have been studied and described from a theoretical point of view in different fields of investigation, a modeling formalization is still missing, and this is the novelty that the present work introduces: the possibility to incorporate all these different aspects into a model of decision-making. A still very primitive framework has been proposed allowing the integration of non-analytical factors into a coherent frame. We achieve such integration by taking into account qualitative definitions of non-analytical factors that stem from different fields of investigations, and quantifying them within the framework of a generalized Markov Model decision process. In this context, the importance of a modeling approach resides in its capacity to focus on the principal and essential factors involved in a process, in this case of decision-making, concisely and practically describing each of them and meanwhile maintaining an overview on the entire phenomenon. We hope that this initial step could lead to further exploration, and deepen each aspect to increase the model details, such as interaction with others, which some preliminary results have been already founded by the authors.

This study does not have a specific psychological connotation, instead it attempts to integrate current cognitive psychology research with the more varied—and inherently uncertain—outcomes of human decision-making and could contribute to the introduction of new aspects, expanding research in this field, such as self-observation and the ensuing emotion, and the use of unexplicit information in the decision. The psychological (and philosophical) dimensions of awareness were, in turn, deeply investigated by Drigas and Mitsea (2020) in terms of metacognition by stressing the need for a reflexive act in which the decision maker acts as a “third person”, retrospectively evaluating their previous strategies and consequently building up a “tacit knowledge reservoir”. It is not without consequence that the authors insert one of the basic pillars of metacognition, “the internalized knowledge that awakens and drives humans towards independence and the fulfillment of each one’s potential” (Drigas and Mitsea, 2020).

There is a large consensus about the presence of two distinct mechanisms in order to tackle the relationship between information and the decision process that we have called intuitive and analytical; which here have been developed, suitably revisited and extended. First, the presence of the phenomenon of overfitting derives from excessive confidence in the analytical approach, which leads the individual to focus on the details of a specific sample that is part of a much wider "population", losing generalization power and potentially moving towards poor predictions and thus poor decisions. We could imagine a kind of threshold beyond which the logical and analytical approach of collecting and analyzing data becomes disadvantageous. Indeed, beginning to model the singularities of the particular reference set that have no equivalent outside the narrow scope from which the data may prevent considering properties "common" to a certain class of problems. We mathematically formulate this phenomenon by introducing the *forward probability transition* of an analytical individual (see Figure 3B), claiming that after a certain threshold of u_t, a further level of analyticity results as a decrease in the probability of reaching a higher value of awareness. On the other hand, the *forward probability transition* of an intuitive individual claims that the more intuitive the reasoning (the smaller u_t), the bigger the probability of increasing the state is until a given lower bound is reached. This happens because the individual thinks they have access to personal abilities, distinct from cognition, allowing the level of awareness to increase by using a kind of unexplicit acquaintance related to tacit knowledge. Thus, the idea expressed by Pascal's *esprit de finesse*, an effortless ability available to each individual but often unknown, is accounted for by our model. Tacit knowledge is inherently difficult to express, extract or demonstrate with objective data but, despite these setbacks, it could possibly be formalized which is one of the novelties introduced in this work.

The model questions a purely analytical "one-size-fits-all" approach, stressing the importance of considering the uniqueness of each single individual who could in any case autonomously change their habits thanks to the implementation of a kind of self-observation mechanism, and recognizing the effectiveness of their personal and unique repository of tacit knowledge.

Moreover, the specificity of an MDP allows bidirectional vision with a look to the future in the evaluation of the optimal choice at each time instant, and a retrospective reconstruction of the entire sequence of choices and the dynamics of the state enabling different perspectives of observation. Considering time an independent variable it is possible to observe the mechanisms by which the state evolves. The model does not provide a univocal definition of awareness, but rather considers it as the result of a series of processes, as described above, which can allow the individual to retrospectively observe the process that led them to be the person they are today.

In the end, interesting aspects arise from the introduction of emotions in the model. We have started from the consideration that emotions impact an individual's intertemporal choice, modifying perceived utility and leading people to behave in ways that seem to disregard the future, thus sometimes damaging the individual themselves. All these aspects are considered in the definition of weight δ that the DM attributes to future rewards. Typically, when included, emotions are evaluated as "noise" to avoid or minimize. The different point of view proposed in this work claims that emotions do not necessarily hurt an individual's choice, they can be "helpful". This depends on one's level of awareness, which can be considered as strictly related to the ability to manage and integrate emotions in decision-making, and in turn enhance the individual's awareness. From the simulations it is possible to appreciate the validity of the above considerations. Starting from a low initial level of awareness in both analytical and intuitive individuals (Figures 7B, 8B), emotions have a damaging effect. The difference is that in the first case the state irrevocably decreases to minimal one, whereas in the second it stopped at a local value without increasing anymore. The analytical case can be interpreted as the typical idea that emotions disturb choice, but, in our model this is only true when considering low states of awareness. At low states the analytical individual is not able to relate with and manage emotions, and this reflects their state decreasing to zero. In the intuitive case, on the other hand, the state stops increasing due to the appearance of an oscillatory dynamic in choice, where the decisions oscillate from a low to a high value without maintaining a constant trend in time. Emotions create an unstable dynamic that does not permit constant and long-lasting decisions over time, resulting in a stationary state. The interesting aspect is that at high levels of awareness, these behaviors do not manifest, and indeed emotions can exert a beneficial influence (Figures 7B, 8C). This is more evident in the intuitive case, whereas in the analytical case they improve in the transient before the state stabilizes to a constant value. This is an indication that emotions are not necessarily a nuisance in the decision process.

Another relevant result has to do with the mathematical formalization of emotions that resonates with the concept of awareness, typical of oriental traditions as connected to the capacity of living in the "*present moment*", where the individual is focused on the present occurrence of experience without interference from past or anticipated images (Kang and Whittingham, 2010). It is not by chance that the exhortation of "living in the present moment" is shared by diverse philosophies, from monastic Christian (Merton, 2010) to mindfulness techniques (Carpenter et al., 2019). Living in the present moment does not mean to be prey to the search of immediate gratification (that means weighing the future with a negative delta), rather it corresponds to the capacity to give an equal weight to each instant (in our case having a delta that reaches one). This is exactly what happens in our model by increasing time and s_t, thus becoming older and more aware.

Conclusions

This work incorporates essential drivers for human decisions, analyzing their reciprocal relations and influences into a model grounded in the Markov process. From the analytical/intuitive dichotomy to the inclusion of tacit knowledge and the impact of emotions, all the different facets of a decision have been discussed from both a theoretical and a mathematical point of view. Individual awareness emerges from the comparison between habitual strategies and the ones sprung from the addition of an individual self-awareness feedback, and its dynamic nature can be appreciated from an individual's retrospective observation. Moreover, the impact of emotions is re-thought with an explicit dependence on the level of awareness of the individual, so that their conception that emotion is a noise to be filtered is mitigated by the consideration that it is true at low state of awareness, and can thus be enhancing for aware individuals. From an epistemological point of view, our aim was to demonstrate how commonly first sight decisions are taken for granted (resonating in diffuse expressions like "clinical eye"), and cannot be considered as a purely "emotional" process opposing "strictly analytical strategies"; instead, they are the result of a long and largely tacit learning process. This concept was already present in the words of Blaise Pascal nearly 400 years ago but progressively forgotten by specialist literature. Here, we give a proof-of-concept of the possibility to insert this kind of knowledge into a mathematical model alongside the philosophical issues we think this result could help solve, from problems encountered by machine intelligence to facing problems relevant for biomedical applications (Gavrishchaka et al., 2019; Beaulieu-Jones et al., 2021).

The limitations are the obvious and inherent ones in creating a mathematical model of such complex phenomena as human decisions and awareness are. Mathematically, modeling awareness is a herculean task, and the model will inevitably be "sloppy" due to the inability to enclose the immensity of human thought into a few functions that are, at best, a stimulus for a more realistic consideration of decision-making process.

One way to overcome the above limitations could be by testing the model in reality, for example developing surveys and designing experiments that can allow for the collection of estimations of the model's parameters and adapt the model to specific cases. A second crucial step forward in model-understanding is to also consider the presence of interactions among individuals. Taking a cue from some preliminary results obtained by the authors in this direction, there is a plan to investigate the effects on the decision process and awareness evolution introduced by interaction within a network of individuals and the different impacts due to the structure of the relationships.

Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Author contributions

CM, AG, and FB: Conceptualization and writing. CM and FB: Mathematical modeling, simulation, and software development. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This study was financially supported by University of Siena.

Acknowledgments

FB, AG, and CM are grateful to Silvia Tiezzi and Alessandro Innocenti for their fruitful comments on awareness as behavioral economics phenomenon, to Duccio Pianigiani for having provided the significant literature on a philosophical approach to awareness. CM also thanks Mohsen Sariaslani for his awareness. The study was financially supported by University of Siena.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adinolfi, P., and Loia, F. (2021). Intuition as emergence: bridging psychology, philosophy and organizational science. *Front. Psychol.* 12, 787428. doi: 10.3389/fpsyg.2021.787428

Alhamzawi, R., and Ali, H. T. (2018). The Bayesian adaptive lasso regression. *Math. Biosci.* 303, 75–82. doi: 10.1016/j.mbs.2018.06.004

Allinson, C. W., and Hayes, J. (1996). The cognitive style index: a measure of intuition-analysis for organizational research. *J. Manag. Stud.* 33, 119–135. doi: 10.1111/j.1467-6486.1996.tb00801.x

Axelrod, S. D. (2005). Executive growth along the adult development curve. *Consult. Psychol. J. Pract. Res.* 57, 118–125. doi: 10.1037/1065-9293.57.2.118

Beaulieu-Jones, B. K., Yuan, W., Brat, G. A., Beam, A. L., Weber, G., Ruffin, M., et al. (2021). Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digital Med.* 4, 1–6. doi: 10.1038/s41746-021-00426-3

Bechara, A., and Damasio, A. R. (2005). The somatic marker hypothesis: a neural theory of economic. *Games Econ. Behav.* 52, 336–372. doi: 10.1016/j.geb.2004.06.010

Beck, A. T., Baruch, E., Balter, J. M., Steer, R. A., and Warman, D. M. (2004). A new instrument for measuring insight: the beck cognitive insight scale. *Schizophr. Res.* 68, 319–329. doi: 10.1016/S0920-9964(03)00189-0

Bellman, R. E., and Drayfus, S. E. (2015). *Applied Dynamic Programming, Vol. 2050*. Princeton, NJ: Princeton University Press.

Bertsimas, D., and Tsitsiklis, J. (1993). Simulated annealing. *Stat. Sci.* 8, 10–15. doi: 10.1214/ss/1177011077

Bird, A. (2010). Eliminative abduction: examples from medicine. *Stud. History Phil. Sci. A* 41, 345–352. doi: 10.1016/j.shpsa.2010.10.009

Bizzarri, F., and Mocenni, C. (2022). *Awareness*. San Francisco, CA: *Academia Letters*.

Brockmann, E. N., and Anthony, W. P. (2002). Tacit knowledge and strategic decision making. *Group Organization Manag.* 27, 436–455. doi: 10.1177/1059601102238356

Brockmann, E. N., and Anthony, W. P. (1998). The influence of tacit knowledge and collective mind on strategic planning. *J. Manag. Issues*, 10: 204–222.

Carden, J., Jones, R. J., and Passmore, J. (2022). Defining self-awareness in the context of adult development: a systematic literature review. *J. Management Educ.* 46, 140–177. doi: 10.1177/1052562921990065

Carpenter, J. K., Conroy, K., Gomez, A. F., Curren, L. C., and Hofmann, S. G. (2019). The relationship between trait mindfulness and affective symptoms: a meta-analysis of the Five Facet Mindfulness Questionnaire (FFMQ). *Clin. Psychol. Rev.* 74, 101785. doi: 10.1016/j.cpr.2019.101785

Christian, B., and Griffiths, T. (2016). *Algorithms to Live by: The Computer Science of Human Decisions*. London: Macmillan.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: BMJ.

Damasio, A. R. (2012). *Self Comes to Mind: Constructing the Conscious Brain*. New York, NY: Vintage.

Dane, E. (2007). Exploring intuition and its role in managerial decision making. *Acad. Manag. Rev.* 32, 33–54. doi: 10.5465/amr.2007.23 463682

Descartes, R. (2008). *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences*. Oxford: Oxford World's Classics.

Diaconis, P., and Mazur, B. C. (2003). The problem of thinking too much. *Bull Am Acad Arts Sci.* 56, 26–38. Available online at: http://www.jstor.org/stable/3824296

Dierdorff, E. C., and Rubin, R. S. (2015). *We're not Very Self-aware, Especially at Work*. Boston, MA: Harvard Business Review.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46. doi: 10.1111/j.1600-0587.2012.07348.x

Drigas, A., and Mitsea, E. (2020). The 8 pillars of metacognition. *Int. J. Emerg. Technol. Learn.* 15, 162–178. doi: 10.3991/ijet.v15i21.14907

Drigas, A., and Mitsea, E. (2020). 8 Pillars X 8 Layers Model of Metacognition: Educational Strategies, Exercises andTrainings. *Int. J. Online Biomed. Eng.* 17, 115–134. doi: 10.3991/ijoe.v17i08.23563

Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis) information. *Am. J. Med.* 113, 763–765. doi: 10.1016/S0002-9343(02)01473-0

Figueiredo, J. (2021). *Tacit Knowledge, Action, Learning, and Spirits of Science*. San Francisco, CA: Academia Letters.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Gavrishchaka, V., Senyukova, O., and Koepke, M. (2019). Synergy of physics-based reasoning and machine learning in biomedical applications: towards unlimited deep learning with limited data. *Adv. Phy. X* 4, 1582361. doi: 10.1080/23746149.2019.1582361

Giuliani, A. (2018). Sloppy models: why in science too much precision is a curse. *Riv. Filos. Neo Scolast.* 4, 737–749. doi: 10.26350/001050_000079

Halpern, J. Y., and Piermont, E. (2020). "Dynamic awareness," in *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning - Main Track*. p. 476–484. doi: 10.24963/kr.2020/48

Halpern, J. Y., Rong, N., and Saxena, A. (2010). "Mdps with unawareness," in *Proceedings of the Twenty-Sixth Conference on Uncertainty in AI*. p. 228–235. doi: 10.48550/arXiv.1407.7191

Heifetz, A., Meier, M., and Schipper, B. C. (2013). Dynamic unawareness and rationalizable behavior. *Games Econ. Behav.* 81, 50–68. doi: 10.1016/j.geb.2013.04.003

Ho, S. Y., Wong, L., and Goh, W. W. (2020). Avoid oversimplifications in machine learning: going beyond the class-prediction accuracy. *Patterns* 1, 100025. doi: 10.1016/j.patter.2020.100025

Hodgkinson, G. P., Langan-Fox, J., and Sadler-Smith, E. (2008). Intuition: A fundamental bridging construct in the behavioural sciences. *Br. J. Psychol.* 99, 1–27. doi: 10.1348/000712607X216666

Kahneman, D. (2017). *Thinking, fast and slow*. Macmillan.

Kang, C., and Whittingham, K. (2010). Mindfulness: a dialogue between Buddhism and clinical psychology. *Mindfulness.* 1, 161–173. doi: 10.1007/s12671-010-0018-1

Karni, E., and Vierø, M. L. (2017). Awareness of unawareness: a theory of decision making in the face of ignorance. *J. Econ. Theory* 168, 301–328. doi: 10.1016/j.jet.2016.12.011

Krishnan, A., Giuliani, A., and Tomita, M. (2007). Indeterminacy of reverse engineering of gene regulatory networks: the curse of gene elasticity. *PLoS ONE* 2, 562. doi: 10.1371/journal.pone.0000562

Landry, P., Webb, R., and Camerer, C. F. (2021). A neural autopilot theory of habit. *Soc. Sci. Res. Network*. doi: 10.2139/ssrn.3752193

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *Am. Econ. Rev.* 90, 426–432. doi: 10.1257/aer.90.2.426

Merton, T. (2010). *The Silent Life*. Farrar: Straus and Giroux.

Mintzberg, H. (2000). *The Rise and Fall of Strategic Planning*. London: Pearson Education.

Mocenni, C., Montefrancesco, G., and Tiezzi, S. (2019). A model of spontaneous remission from addiction. *Int. J. of Appl. Behav. Econ.* 8, 21–48. doi: 10.4018/IJABE.2019010102

Modica, S., and Rustichini, A. (1999). Unawareness and partitional information structures. *Games Econ. Behav.* 27, 265–298. doi: 10.1006/game.1998.0666

Modica, S., and Rustichini, A. (1994). Awareness and partitional information structures. *Theory Decis.* 37, 107–124. doi: 10.1007/BF01079207

Papaleontiou-Louca, E. (2003). The concept and instruction of metacognition. *Teach. Dev.* 7, 9–30. doi: 10.1080/13664530300200184

Pascal, B. (2012).. *The Thoughts. Kegan, P.C. (English translation)*. Toledo, OH: Veritatis Splendor Publications.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *Int. Conf. Mach. Learn.* 70, 2778–2787. doi: 10.1109/CVPRW.2017.70

Polanyi, M. (2009). *The Tacit Dimension*. Chicago, IL: University of Chicago press.

Prietula, M. J., and Simon, H. A. (1989). The experts in your midst. *Harv. Bus. Rev.* 67, 120–124.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley and Sons.

Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.*, 9, 545–556. doi: 10.1038/nrn2357

Reber, A. S. (1993). *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive*. New York, NY: Oxford University Press.

Sayegh, L., Anthony, W. P., and Perrew,é, P. L. (2004). Managerial decision-making under crisis: the role of emotion in an intuitive decision process. *Hum. Resour. Manag. Rev.* 14, 179–199. doi: 10.1016/j.hrmr.2004.05.002

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social* 44, 695–729. doi: 10.1177/0539018405058216

Soosalu, G., Henwood, S., and Deo, A. (2019). Head, heart, and gut in decision making: development of a multiple brain preference questionnaire. *SAGE Open* 9. doi: 10.1177/2158244019837439

Tayor, F. W. (2004). *Scientific Management*. Abingdon: Routledge.

Transtrum, M. K., Machta, B. B., Brown, K. S., Daniels, B. C., Myers, C. R., and Sethna, J. P. (2015). Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys.* 143, 010901. doi: 10.1063/1.4923066

Vaneechoutte, M. (2000). Experience, awareness and consciousness: suggestions for definitions as offered by an evolutionary approach. *Found. Sci.* 5, 429–456. doi: 10.1023/A:1011371811027

Wessinger, C. M., and Clapham, E. (2009). *Cognitive Neuroscience: An Overview. Encyclopedia of Neuroscience*. Oxford: Academic Press. p. 1117–1122. doi: 10.1016/B978-008045046-9.00299-0

Xie, Y. L., and Kalivas, J. H. (1997). Evaluation of principal component selection methods to form a global prediction model by principal component regression. *Anal. Chim. Acta* 348, 19–27. doi: 10.1016/S0003-2670(97)00035-4

TYPE Original Research
PUBLISHED 13 December 2022
DOI 10.3389/fpsyg.2022.1020770

OPEN ACCESS

EDITED BY
Anna M Borghi,
Sapienza University of Rome, Italy

REVIEWED BY
Frank Zenker,
Boğaziçi University,
Turkey
Kimmo Eriksson,
Mälardalen University,
Sweden

*CORRESPONDENCE
Ross Jacobucci
rjacobuc@nd.edu

SPECIALTY SECTION
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

RECEIVED 16 August 2022
ACCEPTED 18 October 2022
PUBLISHED 13 December 2022

CITATION
Jacobucci R (2022) A critique of using the labels *confirmatory* and *exploratory* in modern psychological research. *Front. Psychol.* 13:1020770.
doi: 10.3389/fpsyg.2022.1020770

COPYRIGHT
© 2022 Jacobucci. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A critique of using the labels *confirmatory* and *exploratory* in modern psychological research

Ross Jacobucci*

Department of Psychology, University of Notre Dame, Notre Dame, IN, United States

Psychological science is experiencing a rise in the application of complex statistical models and, simultaneously, a renewed focus on applying research in a confirmatory manner. This presents a fundamental conflict for psychological researchers as more complex forms of modeling necessarily eschew as stringent of theoretical constraints. In this paper, I argue that this is less of a conflict, and more a result of a continued adherence to applying the overly simplistic labels of *exploratory* and *confirmatory*. These terms mask a distinction between exploratory/confirmatory *research practices* and *modeling*. Further, while many researchers recognize that this dichotomous distinction is better represented as a continuum, this only creates additional problems. Finally, I argue that while a focus on preregistration helps clarify the distinction, psychological research would be better off replacing the terms exploratory and confirmatory with additional levels of detail regarding the goals of the study, modeling details, and scientific method.

KEYWORDS

exploratory research, confirmatory research, big data, machine learning, philosophy of science

Introduction

Psychology has seen a renewed interest in the application of confirmatory research (e.g., Wagenmakers et al., 2012; Scheel et al., 2020), spurred on by the so-called *replication crisis* (e.g., Maxwell et al., 2015). Of the many factors which play a part in the replication crisis, a consistent theme is that researchers take liberty with several steps in proceeding from concept formation to deriving statistical predictions (Scheel et al., 2020). Characteristically, this flexibility in conducting confirmatory research (i.e., researcher degrees of freedom; Simmons et al., 2011) resulted in the label of "exploratory research" or "wonky stats" (Goldacre, 2009; Wagenmakers et al., 2012) along with "non-confirmatory" research (Scheel et al., 2020) for studies that do not follow a strict confirmatory protocol.

For much of the history of psychological research, just using the terms confirmatory and exploratory were justified, as the majority of psychological research was concerned with theory appraisal, explanation, and first generating hypotheses or theory, followed by testing the predictions deriving from it (Hypothetico-Deductive method; detailed further below). However, the advent of big data, and the corresponding algorithms, has shifted a subset of research, resulting in study procedures and aims being less likely to follow default procedures. Fundamentally, psychological research is increasingly incorporating elements that could be construed as exploratory, with this reflecting a natural progression as models grow increasingly complex. Applying the relatively simplistic labels of exploratory and confirmatory to research papers comes with a number of drawbacks. Exploratory research is often devalued,[1] regardless of the context in which this takes place. As a result, to describe their research paper as confirmatory, researchers often hide any degree of uncertainty as to the theoretical foundations to their research, fail to report any model modifications made, among many other practices (e.g., Simmons et al., 2011; Gelman and Loken, 2014). Further, researchers often use the term exploratory without regard for the context in which the research takes place: exploratory practices in experimental studies (e.g., Franklin, 2005) are quite distinct from atheoretical studies that apply machine learning to large datasets.

I make the case in this paper that using the labels of exploratory and confirmatory applied to studies as a whole, or aims within a study, come with four primary limitations:

- Conflates research practices and modeling.
- The labels are often used in global ways, masking local decisions.
- Are better represented as a continuum as opposed to discrete labels, but results in arbitrary placements.
- Often serve as proxies for more descriptive terms or procedures.

Failing to detail the motivations behind the study, reasoning underlying the procedures, and the large number of decisions made regarding the data and models severely hinders the study's impact in building a cumulative body of research. While a large number of papers have advocated for increased reporting of *how* a study was conducted (e.g., Depaoli and Van de Schoot, 2017; Aczel et al., 2020), the focus of this paper is on *why* a study was conducted in the manner reported. Prior to discussing these four primary limitations I provide background on the ways in which the terms confirmatory and exploratory have been depicted in psychological research. This is followed by a set of recommendations for moving beyond simply placing studies or study aims into the overly simplistic boxes of exploratory or confirmatory. Instead, I advocate for providing detail on how replication/generalizability was addressed, the form of reasoning used, and orienting the research with respect to explanation, prediction, or description, as well as theory generation, development or appraisal.

What do the terms *exploratory* and *confirmatory* mean?

Confirmatory research is a hallmark of science. Stating hypotheses, running a study or experiment to test these hypotheses, and then either finding enough evidence to support or fail to support the hypothesis, is an efficient and important cornerstone to this practice. A common view of what constitutes confirmatory research is a series of *a priori* hypotheses followed by developing a research design (experimental in most cases) to test the hypotheses, gathering data, analysis, and concluded with deductive inference (Jaeger and Halliday, 1998). Hypotheses can only be refuted or not refuted, typically assessed using parametric statistics and p-values (Null Hypothesis Significance Testing). One prevailing distinction between confirmatory and exploratory research is in the tradeoff between Type I and Type II errors, with confirmatory research favoring low Type I errors, and exploratory research preferring low Type II error (Snedecor and Cochran, 1980).[2] This can also result in different critical alpha levels being uses in assessing p-values, with exploratory research allowing more liberal conclusions (Jaeger and Halliday, 1998).

The most detailed counter to confirmatory research is exploratory data analysis (EDA). While the term EDA has been used in many ways, referring to both research practices and data analysis, the most common meaning refers to the seminal work of Tukey (1977). Tukey (1977) almost exclusively focuses on the use of data visualization to carry out EDA. This can be used in the case of simply exploring data in examining single variables, assessing the assumptions of relatively simple models such as linear regression, or examining the concordance of the actual data and model implied data in complex models (Gelman, 2004).

Going beyond EDA, the definition of exploratory research is most often defined in terms of what it is not, as opposed to what it is. While confirmatory research involves testing hypotheses, exploratory research involves hypothesis generation: "Explicit hypotheses tested with confirmatory research usually do not spring from an intellectual void but instead are gained through exploratory research" (p. S64; Jaeger and Halliday, 1998). This is in line with Good's (1983) description of EDA as a mechanism to deepen a theory, by looking for holes (residuals).[3] It is not that exploratory research is devoid of hypotheses, but instead that the hypotheses are often relatively vague and may evolve over the course of experiments or analyses (Kimmelman et al., 2014).

[1] An extreme example is Lindsay (2015, p. 1828) "Personally, I aim never again to submit for publication a report of a study that was not preregistered. Exploratory work has great and essential value but is typically not appropriate for publication."

[2] This can be equivalently described as exploratory research giving preference to sensitivity, while confirmatory are more concerned with specificity (Kimmelman et al., 2014).

[3] More contemporary research is more likely to utilize additional models or algorithms that have different functional forms or assumptions, capitalizing on developments in computational power (see Lee et al., 2019).

Data types

One consistent distinction is in the types of data aligned with each type of modeling, with confirmatory data analysis (CDA) using mostly experimental data, while EDA typically uses observational data (Good, 1983). Though CDA involves specification of hypotheses to directly inform data collection, EDA is often conducted on data that were collected informally, or secondary data analysis. Further, the terms exploratory and confirmatory are almost exclusively contrasted with respect to more traditional data types, with much less description with respect to big data. However, psychological research is increasingly collecting and analyzing new types of data, such as from magnetic resonance imaging, various types of text data, actigraphy data, to name just a few. These new data types present many opportunities for novel types of hypotheses, additional modeling flexibility, along with some challenges to traditional ways of thinking about exploratory and confirmatory research. For instance, can something be exploratory if there is no available confirmatory counterpart? As an example, datasets with more variables than samples require the use of methods such as ridge regression to overcome computational difficulties faced by ordinary least squares. This severely limits the potential for imparting theory into the analysis, as algorithm constraints, not *a priori* theoretical motivations, are required to reduce the dimensionality of the model. In a larger sense, new data types have the potential to further the distinction between theory, variable selection, and the actual models that are tested, further muddying a study's labeling of confirmatory/exploratory.

To describe this further, we can address the following question: what does confirmatory research look like in the context of text responses? Text data is not unique in the respect of having very few (if no) studies that are confirmatory in nature, as it is more a characteristic of studies that utilize high-dimensional data, as detailed somewhat previously with P > N datasets. Traditional regression models allow researchers to impart theory in the variables used, sequence of models tested, the use of constraints (such as in the form of no relationship constraints in SEM), or the use of interaction terms or testing mediation models, among others. Each of these theoretical characteristics of regression models do not have an analog in text algorithms, or if they do exist, require a great deal of simplification to make the results interpretable (for instance using dictionary-based approaches such as LIWC (Pennebaker et al., 2001) which are based on *a priori* created dictionaries of words, which have clear drawbacks [e.g., Garten et al., 2018]).

In contrast to the use of dictionaries, text data is most often analyzed using relatively complex latent variable models such as latent dirichlet allocation (LDA; Blei et al., 2003) or latent semantic analysis (LSA; Deerwester et al., 1990). This allows researchers to pose hypotheses related to the existence of latent topics common to participants text responses. Relative to psychology research, these are similar in structure to mixture models (LDA) or factor analysis (LSA). However, in contrast to a method such as factor analysis, neither LDA or LSA allow for theory-based constraints as in the form of specifying specific factor loadings. Further, researchers can move beyond the use of LDA and LSA and model the sequence in which words are used with a host of neural network models (e.g., Mikolov et al., 2013). This requires larger amounts of data, but affords modeling the text in a way that is more in line with the way that the words were produced. Using neural networks to model text represents an extremely complex form of modeling, allowing very little theoretical input. In summary, the complexity of text data restricts the degree of theory that can be imposed to into the statistical models, thus meaning analyzing data of this type would exclusively be referred to as exploratory. Ultimately, new types of data necessarily are paired with less theoretical foundation, thus lending themselves to modeling with more to induction than to theory testing (or perhaps more clearly to discovery as opposed to justification, i.e., Reichenbach, 1938; Howard, 2006).

Preregistration

In assessing research articles, readers are required to place trust in the authors that the sequence of procedures that was stated in the article mimics what was done in practice. This trust has come in to question spurred on by the replication crisis (e.g., Morawski, 2019), prompting methods such as preregistration to be proposed as a remedy, which has quickly become popular in psychological research (Simmons et al., 2021). Preregistration allows researchers to state the temporal sequence to hypotheses and analyses, thereby increasing the credibility to the statements made in published research.

Using preregistration as a form of validation to better delineate confirmation and exploration may best be summarized in the following: "First, preregistration provides a clear distinction between confirmatory research that uses data to test hypotheses and exploratory research that uses data to generate hypotheses. Mistaking exploratory results for confirmatory tests leads to misplaced confidence in the replicability of reported results" (Nosek and Lindsay, 2018). Additionally, in differentiating between exploration and confirmation, one can go beyond the distinction of whether hypotheses were stated *a priori*, and delineate whether analyses were planned (confirmation) or unplanned (exploration), mimicking the distinction made above between practices and modeling. This does not have to be the case, but is often equated (Nosek et al., 2019).

Conflating practices and modeling

In the above characterizations of exploratory and confirmatory, there is a conflation between confirmatory vs. exploratory *research practices*, and confirmatory vs. exploratory *modeling* or *data analysis*. While exploratory can refer to EDA, it can simultaneously be taken to mean not specifying *a priori* hypotheses. On the flip side, the term confirmatory can refer to preregistering hypotheses to ensure that they were in fact stated prior to data collection and analysis, or in the case of larger datasets, the use of confirmatory factor analysis to test a hypothesis regarding latent variables. Below, I provide further distinctions of each.

Research practices

More recently, almost a consensus has occurred that researchers can no longer be trusted to accurately detail the steps they took in conducting their study (Moore, 2016; van't Veer and Giner-Sorolla, 2016; Nosek et al., 2018). By not preregistering aspects of study design or analysis, researchers are afforded a degree of flexibility that can compromise the veracity of resultant conclusions. This has been referred to as researcher degrees of freedom (Simmons et al., 2011; Gelman and Loken, 2014), and often manifests itself as multiple comparisons, or "fishing," and then reporting the best result. Instead, researchers should preregister the study design and analysis plan, among other components of their study. This offers a safeguard against the reporting of exploratory results as if they were confirmatory, namely saying that the hypotheses were stated prior to the analysis results, not the other way around.

One complication in labeling research practices as exploratory is the blurry line between what is termed exploratory, and what is considered questionable research practices (QRPs; Simmons et al., 2011; John et al., 2012). For instance, exploratory is characterized as "where the hypothesis is found in the data" in Wagenmakers et al. (2012), while a number of QRPs revolve around whether descriptions of the results match the order in which the study took place, possibly best exemplified by "claiming to have predicted an unexpected finding (John et al., 2012). Underlying this distinction is the motivation behind the research practice, which can only be assessed through reporting. Thus, without detailed reporting standards, it is easy to conflate exploratory research with QRPs, thus further disadvantaging those that are truly conducting exploratory research.

This leads into a second component of confirmatory research: the data used to confirm hypotheses (test set) must be separate from the data used to generate the hypotheses (train set). Given the small samples sizes inherent in psychological research (i.e., Etz and Vandekerckhove, 2016), this strategy can rarely be fulfilled in practice. This is often referred to as cross-validation and has seen an upsurge of interest in psychology (Koul et al., 2018; de Rooij and Weeda, 2020). To clear up one point of confusion with regard to the term cross-validation, I first need to distinguish two similar, but separate strategies. We can term the strategy of splitting the sample into two separate datasets the validation set approach (e.g., James et al., 2013; Harrell, 2015), also referred to as the *Learn then Test* paradigm (McArdle, 2012). This strategy is often recommended in both psychological and machine learning research but is rarely used due to requiring large initial sample sizes. A second cross-validation strategy involves only using one dataset but repeating the process of splitting the sample into train and test sets, and selecting different subsets of the data for each split. This form of resampling is most commonly conducted using *k* separate partitions of the data (*k*-fold cross-validation) or the repeated use of bootstrap samples. In both *k*-fold and bootstrap sampling, the part of the sample not used to train the model is used to test the fixed model, allowing for a less biased assessment of model fit. It is important to note that most papers that describe cross-validation as a viable strategy to separate exploratory from confirmatory research (e.g., Behrens, 1997; Wagenmakers et al., 2012; Fife and Rodgers, 2022) are referring to the validation set approach rather than *k*-fold cross-validation or bootstrap sampling. While Haig (2005) describes internal validation procedures, such as the bootstrap or *k*-fold cross-validation, as confirmatory procedures, this statement rests on the assumption that the analytic tool being applied has a low propensity to overfit the data, thus the bootstrapped assessment of fit is close to unbiased. In machine learning, the whole sample fit can be extremely positively biased (for example, see Jacobucci et al., 2021), thus internal validation is required (and ideally external validation) to derive a realistic assessment of initial fit, as the within sample fit is often unworthy of examination.

Modeling

Specific statistical methods are often labeled as being exploratory or confirmatory, which typically involves the degree of theoretical specification that a model/algorithm affords. On the exploratory side of the spectrum is EDA, machine learning, and exploratory factor analysis, while linear regression, confirmatory factor analysis, and ANOVA are often characterized as being confirmatory. The distinction is often based on the degree of constraints that a statistical method imposes on the data, with these constraints (e.g., linearity or setting specific relationships to be zero) aligning with specific theoretical foundations. For instance, structural equation modeling (of which regression can be seen as a subset of) allows researchers a large degree of flexibility in the type of relationships that can be specified based on theory, and equally as important, which relationships are specified to be non-existent. Further, from a realist perspective, latent variables are defined as real entities, which is difficult to justify from an exploratory (atheoretical) perspective (Rigdon, 2016). In contrast, machine-learning algorithms are often described as atheoretical or exploratory, which can mainly be ascribed to the lack of opportunity afforded researchers to test or impose specific relationships. Instead, relationships are learned from the data.

This may be best exemplified in the context of confirmatory factor analysis. As an example, one can imagine assessing depression, stress, and anxiety with the Depression Anxiety and Stress Scale (DASS-21; Lovibond and Lovibond, 1995). With this, three latent variables would be posed, and in an ideal scenario, a researcher has a fully specified factor model, which mainly involves assigning which observed variables load on which latent variables. A dilemma is faced in the common result of the fit indices indicating some degree of non-optimal fit, which could either be evidenced consistently or inconsistently across multiple indices. Often, this results in researchers searching among the modification indices, making small tweaks here and there to residual covariances, which are typically not reported in the manuscript (Hermida, 2015). Alternatively, if the fit indices are too far away from "good" fit to be salvaged by a handful of *post-hoc* modifications, researchers could return to the "exploratory" phase, often using exploratory factor analysis to reassess either the number of latent variables or which observed variables require cross-loadings. This is one of the few options afforded researchers in this position, as more traditional visualization tools aren't designed to address lack of misfit indicated by fit indices, and alternative methods such as the use of modification indices have a more general negative reputation (e.g., see MacCallum et al., 1992).

However, newer types of statistical models, particularly those associated with machine learning, are less likely to allow for constraints based on theoretical justification, instead falling more in line with the "throw everything in" approach to data analysis. This is far too simple of a characterization, but one that is generally proffered around by researchers averse to the use of machine learnings methods. The danger in making distinctions such as this is that nothing about the statistics or math is inherently exploratory. Statistical methods are just that, statistical methods. It is how one uses these methods that make them either confirmatory or exploratory. Further, affixing the term confirmatory to a specific statistical method can obfuscate the lack of strong theory (Lilienfeld and Pinto, 2015), giving the researchers a false sense of confidence to the degree of theory imparted in the study[4].

In examining the labeling conventions based on the degree of constraints the specific statistical methods afford, one quickly runs into contradictions. For instance, EFA actually makes a number of relatively restrictive assumptions, namely that a reflective, not formative model is most appropriate, the relationships are linear, local independence, and researchers can specifically test a hypothesis regarding the number of factors. On the other hand, machine learning algorithms can be used in ways to assess theoretical statements, such as the existence of interactions and/or nonlinearity, fit a linear regression model in the presence of collinearity (using ridge regression), and test new forms of hypotheses (detailed later). Further, the degree of constraint placed on the model or number of parameters does not always align with labeling conventions. For example, lasso regression is often labeled as machine learning despite often having fewer parameters than linear regression. The takeaway point in this discussion is that it is seldom justified to affix the labels of confirmatory or exploratory to specific statistical methods, as almost any method can be used in a confirmatory or exploratory manner (for a similar argument, see McArdle, 1996).

Consequences

A consequence of labeling the preponderance of methods available as confirmatory induces a feeling of guilt when researchers may not have a concrete hypothesis (McArdle, 2012). Instead, researchers skirt the issue by hiding the exploratory nature of the analysis, and only reporting the best fitting model, or the results and conclusions the arrived at after a considerable degree of fiddling (such as using modification indices without reporting) with the data and models, thus confirming the need for preregistration. Further, the devaluation of exploratory methods/questions (i.e., as exploratory) limits transparent theory generation and imply that researchers should always magically have a rigorous hypothesis to test.

A large percentage of modern research does not fit neatly into the above descriptions of exploratory and confirmatory research for a number of reasons. Further, the descriptions of confirmatory research have seen little application to more recent, complex psychological research studies, thus resulting in a large portion of recent, complex research being labeled as exploratory, despite containing multiple theoretical aspects.

The confusion surrounding the distinction between the terms *confirmatory* and *exploratory* is mainly due to a conflation of two separate questions:

1. How much theory is imparted into various aspects of the study/analysis?
2. What steps have been taken to ensure replicability or generalizability?

Whereas the majority of machine learning studies treat these questions as completely separate (especially with respect to point #2 and the use of cross-validation), as well as being relatively straightforward to answer, these questions are evaluated on a single dimension in much of psychological research. Most often, the recommendations made to address the replication crisis comprise both questions, such as advocating for more concrete theories (e.g., Mansell and Huddy, 2018), not conducting exploratory statistics, and preregistration.

One dimension that specifically muddies the distinction between both questions is the lack of reporting characteristics of many research articles. If researchers do not report what models were tested, it is impossible to determine a clear answer to #1, thus altering what steps should be made for answering #2. As a typical example, if researchers only report a single CFA model that fits well and do not report steps taken conducting EFA, various modifications made to the CFA that were based on modification indices, among others, then a false sense of confidence would be placed into the authors answer to questions #2 by reporting various fit indices that have strong simulation evidence for their ability to assess model fit. This is further discussed below with respect to preregistration.

Part of the confusion regarding the distinction between exploratory and confirmatory concerns the term "hypothesis."

[4] This is why some researchers have proposed replacing the term confirmatory factor analysis with structural factor analysis to better denote that the method places structural constraints on the relationships in the data and is not inherently confirmatory (McArdle, 1996).

Most accounts describe a hypothesis as a specific, well-formulated statement. Part of the motivation for this may stem from philosophy of science's fixation on hard sciences, such as physics, where general laws can take mathematical forms. In reality, particularly in psychological research, a hypothesis more often "is nothing but an ebullition of alternative ideas and a pure emotion – consuming speculative curiosity about a certain aspect of the world" (Cattell, 1966). A further problem with the term hypothesis is its generalization from an introductory statistics formulation (i.e., H_0 = no effect) to complex theoretical formulations. A hypothesis taking the form of a single sentence necessarily denotes some form of reductionism from theory, while a hypothesis matching the degree of theoretical complexity would require, at the very least, a paragraph of formulation. Even with the more recent calls to match theory with mathematical structures in the areas of computational modeling (e.g., Fried, 2020), the degree of complexity necessitates some degree of simplification (DeYoung and Krueger, 2020). This critique of how hypotheses are often specified has consequences for determining the degree of theoretical foundation for studies, as vague hypotheses leave considerable leeway in data analysis.

Global versus local

In psychological research, the labels of confirmatory/exploratory have been applied to entire studies, or specific aims/hypotheses within a study. Below, I make the case that applying these labels at the global level can mask inconsistencies in the theoretical rationale for decisions at the local level. The number of local decisions that require theoretical justification to follow a truly confirmatory protocol grows almost exponentially as the size of data and number of algorithms considered grows. This can take place with respect to both modeling and the levels of analysis.

Modeling details

While it is not feasible to describe all the possible aspects that go into a research study, I provide a number of dimensions in Table 1 that are often characteristic of using more complex statistical algorithms, with further detail on what these components look like when based on theory or are atheoretical. The purpose in detailing several dimensions inherent in statistical modeling is to point out how just describing a study as confirmatory or exploratory gives very limited insight into the level (or lack thereof) of theory inherent in each analysis decision.

Other researchers have acknowledged the inherent complexity in modern modeling, while advocating for incorporating both preregistration for detailing decisions gone into confirmatory models, along with postregistration for steps taken in conducting follow up, exploratory analyses (Lee et al., 2019). However, this presumes that despite acknowledging that the confirmatory modeling step has a large number of decisions to be made, and flexibility with regard to their choices, the researchers are able to somewhat confidently decide among this myriad of options to formulate the preregistration plan. In contrast, I am advocating for acknowledging which aspects of the modeling procedure are set based on theory, and which there is some degree of uncertainty. In the end, both perspectives could have the exact same outcome, in detailing a preregistration plan with acknowledgement of certain aspects that are tested in the data.

Note that the term hypothesis is not provided in the above table. This is done due to the inherent complexity to modern hypotheses, as they rarely take either a purely theoretical form outside of experimental contexts, or a purely atheoretical form. Instead, I view it as more fruitful to focus on adding additional detail regarding the aspects detailed in Table 1, thus being more concrete in translating hypotheses to aspects detailed in Table 1 and the sections below.

Level of analysis

While larger datasets better afford the fitting of more complex models,[5] there is not a one-to-one relationship. In fact, larger datasets afford more flexibility in the types of models fit, which can all exist at similar levels of abstraction, or exist across levels. I follow the hierarchy put forth in Kellen (2019; Figure 1), which is based on Suppes (1966), with further elaboration based on this paper's premise.

In contrast to Kellen (2019), a number of changes were made. First, given that this paper's focus is not on experimental research, I label this level study design. Further, I partition this based on the degree of theory that went into study design. This is to account for studies that have a theoretical rationale for every variable assessed (fully theoretical), those that have a rationale for a subset of variables (partially theoretical), and those in which data was not directly collected by the researcher, through openly available datasets or other mechanisms. Within this dimension I wish to acknowledge that a majority of modern research studies collect a large number of variables with the aim to use them for several publications.

The second dimension is the data model, which entails the translation of the raw data to that which is used by the modeling

[5] I define a model not in the statistical sense, where model is often described based on constraints imposed on probability distributions, yielding distinctions such as exploratory factor analysis being a model while principal components analysis is not (e.g., Kasper and Ünlü, 2013). Further, I contrast this definition with formal models (e.g., Smaldino, 2017; also termed computational models, see Robinaugh et al., 2020), where precise statements are made in the form of equations relating phenomenon. Instead, I use a more general definition, following Bailer-Jones (2009): "a model is an interpretive description of a phenomenon which facilitates access to that phenomenon." There is considerable leeway with the phrase "interpretive description," which leads us to further describe proposed hierarchies to types of models.

TABLE 1 Decomposition of multiple study components as to whether the decisions are theory based or non-theory based.

	Theory based	Non-theory based
Algorithm	Algorithm inclusion based on hypothesized relationships in data.	Algorithm inclusion based on convenience or maximizing prediction.
Hyperparameters	Set to be single values or a small set.	Based on software defaults or test a wide range.
Variable inclusion	Each variable is justified.	Variables are chosen based on convenience.
Functional form	Each specified functional form (e.g., linear) should have a theoretical meaning/rationale	Flexibility is inherent in the algorithm to fit a range of functional forms for each relationship
Variable importance/strength	All or a subset of relationships are specified.	Using Ensembles to derive variable importance.
Model choice	Prefer Parsimony.	Prefer Best Fit.

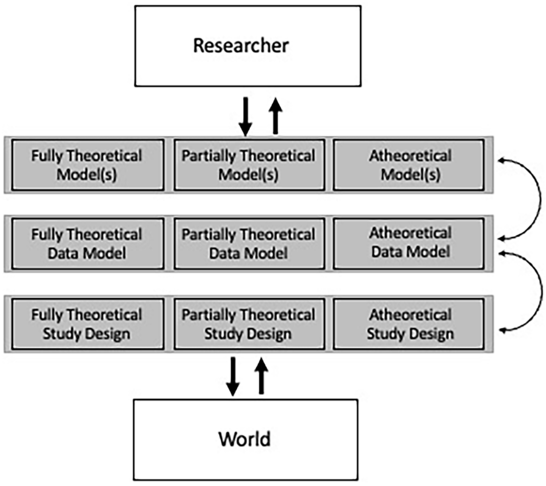

Researcher
Fully Theoretical Model(s)
Partially Theoretical Model(s)
Atheoretical Model(s)
Fully Theoretical Data Model
Partially Theoretical Data Model
Atheoretical Data Model
Fully Theoretical Study Design
Partially Theoretical Study Design
Atheoretical Study Design
World

FIGURE 1
Differentiating various levels of study components based on Suppes (1966) and Kellen (2019).

procedures. In this, raw variables could be used directly or through summaries of sets of variables depending on the theory, or raw data could be transformed by the models through data-driven dimension reduction. In reality, most research would fall into the partially theoretical box, as many studies only use a small number of variables which are based on summaries, or studies directly specify the use of a subset of variables, while others have a small number of variables of interest, but a multitude of additional variables are “tested” in the form of covariates.

The final dimension concerns the degree of theoretical specification in the modeling stage. Researchers have considerable flexibility in translating a theoretical model to something that conforms to an actual dataset. While this is often criticized (e.g., Yarkoni, 2020), there remains considerable utility to working models that can serve as analogies (Bartha, 2013). In psychological research this may be best exemplified in the debate surrounding the field’s understanding of psychological disorders, with reductionism to neurobiological explanations having historical favoring, but more recent pushback from the network analysis research literature (e.g., Borsboom et al., 2019). These debates seemingly mirror those in other fields, with some sides arguing that including high levels of detail can shed light on theoretical aspects that are under-developed (Fried, 2020), while others see the complexity of real data far outmatching even the most detailed computational model (DeYoung and Krueger, 2020), among many other distinctions.

To provide an example that more concretely distinguishes each level and the degree of theory, we can use a moderation model as an example. Notably, simple moderation models almost always entail theoretical specification of each path, resulting in a path diagram (fully theoretical model). However, there is often a discrepancy between the path diagram and the statistical model used (data model), with a multilevel model adhering closer to the path diagram than the regression model with cross-product terms that is typically fit (Yuan et al., 2014). Finally, while the path diagram contains the names of the variables, researchers have flexibility in whether individual items, summed scores, or factor scores are used to represent each variable.

Both “fully theoretical models” and “atheoretical models” are the subject of most description, possibly best exemplified by mediation models and machine learning algorithms such as neural networks, respectively. However, one could argue that the majority of data models do not fall at either extreme, as most “confirmatory” models have at least parts of the model that were not described in the hypotheses or other parts of the theory formulation. Beyond this, there exist a host of statistical methods that facilitate partially theoretical models. An example is mixture models that are combined with other models, such as growth mixture models (e.g., Ram and Grimm, 2009). In this, a latent growth curve model is specified based on theory, then latent classes are estimated that result in fundamentally different growth trajectories across the classes. This latter model component is not directly based on theory, otherwise researchers could specify a multiple group growth model where the heterogeneity to the growth trajectories is based on observed, not latent, groups.

Atheoretical modeling would take the form of specifying many potential algorithms/models, each of which contain varying degrees of interpretation and propensity to fit the data. One caveat with respect to atheoretical modeling is the common scenario, brought about by increased use of machine learning, where researchers specify a number of algorithms, with the conclusions about the best fitting model having theoretical consequences. This is often conducted in machine learning research, where a linear model fitting better or equal that of a neural network model would lead to the conclusion that linear relationships are sufficient to explain the relations between predictors and outcome.[6]

Part of the goal in detailing the complexity to modern modeling is to encourage researchers to test models/algorithms at varying levels of flexibility. In too much research, researchers pose a model at one level of complexity. The problem with this is that the severity of the hypothesis test is minimal (for instance, see Mayo, 1996). As an example, a large body of clinical applications of machine learning only test a single machine-learning algorithm. The severity of this assessment is significantly bolstered by not just showing that the machine learning algorithm fits the data well, but that it fits the data significantly better than a linear model. This is in contrast to other forms of modeling that are more well established, such as latent growth curve modeling. In this, specifying only a quadratic growth model without assessing the improvement in fit over a linear (or other simpler form) growth model would receive swift criticism.

Lastly, I view the relationship between the theoretical model and data model as underdeveloped in most psychological research (see Ledgerwood, 2018 for similar arguments), which is mainly facilitated by a lack of detail regarding the theoretical model, which is possibly most clearly seen in most studies defaulting to the use of summed scores (of which can often be difficult to describe at a conceptual level, McNeish and Wolf, 2020). While this makes sense if a fully theoretical model is posed that depicts relations between latent constructs, this makes far less sense if the model is less than fully theoretical. As an example, network models pose direct relationships between symptoms, which can often be directly assessed in individual questions, while factor models pose that the latent variables represent coherent summaries of the individual items.

While there is a strong correlation between the descriptors theoretical/atheoretical and confirmatory/exploratory, the important piece is that most studies are multidimensional in nature, and each component deserves detail with respect to the degree of theory imparted. Given that the terms confirmatory/exploratory are most often used to describe studies, I believe that these terms should be replaced with theoretical/atheoretical to denote local details of a study.

The exploratory–confirmatory gradient

While the labels exploratory and confirmatory are often ascribed in a dichotomous fashion, a large number of researchers have more appropriately seen them as a continuum (e.g., Scheel et al., 2020; Fife and Rodgers, 2022). However, much less detail has been provided on how one identifies where on this continuum a research study falls, let alone individual aspects of a research study. Outside of the label of "rough CDA" to describe research that is mostly confirmatory but also acknowledges some aspects were derived from the data (Tukey, 1977; Fife and Rodgers, 2022), almost no detail has been provided to describe research more accurately. As an example of what this could look like, labels are placed along the continuum from exploratory to confirmatory in Figure 2.

This is in no way meant to be comprehensive, but instead to depict how a select set of scientific practices would likely fall in terms of exploration and confirmation. Most psychological research likely falls in the middle right of the gradient, with some degree of theoretical specification, but stopping short of making specific statements. The far-right hand side of confirmatory corresponds to what Meehl refers to as strong theory, where specific point predictions are made (see bottom of Meehl, 1997, p. 407). The most common conceptualization of exploratory research would fall on the far left of the gradient, where any form of theoretical specification or hypothesis is eschewed in favor of atheoretical modeling.

In this, it is clear to see that the placement of each phrase is relatively arbitrary and could be up for debate. Further, it is easy to imagine research scenarios that have multiple aspects of the design or analysis that occupy different placements on the continuum. In fact, this one-dimensional continuum is only sufficient in the simplest of psychological studies, whereas most modern psychological research entails a large number of decisions, each of which can include varying degrees of theoretical specification (as discussed above). For instance, covariates in a regression could each have been selected based on theory, however, the regression weights were not constrained based on prior research. While this latter specification could seem rather severe, researchers have options such as this to impart strong theoretical ideas (see McArdle, 1996 for further elaboration). Fried (2020) is a more recent discussion of weak versus strong theory, a distinction that mimics the above contrast between exploratory and confirmatory. One could argue that these seemingly parallel lines of contrast are really one and of the same, with researchers hiding behind the "confirmatory" nature of their study to mask what is in reality quite weak theoretical specification.

In the above gradient, there is a clear hierarchical relationship between some of the practices. For instance, detailing which variables are included comes before statements can be made as to which of these variables are likely to have significant effects, and which are not. This further complicates the use of overarching statements of confirmatory or exploratory about the research paper, as the research practices used in the study represent varying points on the continuum.

Exploratory/confirmatory as proxies

Applying the label confirmatory to describe studies or aims within a study has a significant overlap with whether (1) the study

[6] Its not nearly this simple, as the measurement of predictors (Jacobson et al., 2021), sample size, among a host of additional factors can influence the comparative fit.

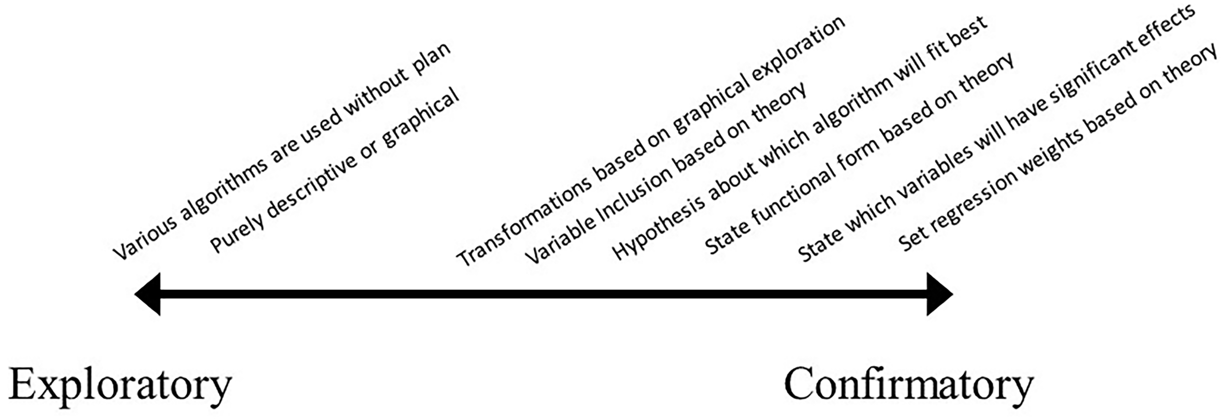

Various algorithms are used without plan
Purely descriptive or graphical
Transformations based on graphical exploration
Variable Inclusion based on theory
Hypothesis about which algorithm will fit best
State functional form based on theory
State which variables will have significant effects
Set regression weights based on theory
Exploratory
Confirmatory

FIGURE 2
A simplified exploratory–confirmatory gradient. Note that this mainly relates to data analysis, not research practices.

adheres to a Hypothetico-Deductive method, (2) the aims are explanatory, and (3) the study is primarily concerned with theory appraisal. Studies or aims that deviate from these procedures often are required to be labeled as exploratory as they likely follow the inductive process, are descriptive or predictive in nature, or concerned with theory generation. Finally, the terms exploratory/confirmatory have been codified in a sense and replaced by preregistration. We elaborate on each of these dimensions below.

Method of science

The contrast between exploratory and confirmatory research masks a more fundamental distinction between forms of method, namely between hypothetico-deductive (HD), inductive, and abductive reasoning. I define these terms as (see Fidler et al., 2018 or Haig, 2014 for more detail):

- Inductive: Moving from the specific to the more general. In research, moving from specific observations based on data to the generation of larger theories or principles.
- Hypothetico-Deductive: Moving from general to the more specific. In research, this is generating hypotheses or theory and test the predictions deriving from it.
- Abductive: Commonly referred to as inference to the best explanation. This involves reasoning about hypotheses, models, and theories to explain relevant facts (see Haig, 2005).

The key distinction between the above is whether hypotheses come prior to the data analysis. In contrast to the hypothetico-deductive method, both inductive and abductive can be seen as reasoning from observation (data). While inductive reasoning combines the creation and justification of theories from observation (Haig, 2020), abduction involves the explanation of empirical relations identified in the data through inference to underlying causes.

The distinction of whether theoretical specification/justification comes prior to or after observation mimics prior discussions on distinguishing between exploration and confirmation. Further, the preference for confirmation is mirrored by HD being the most common method used in scientific research (e.g., Mulkay and Gilbert, 1981; Sovacool, 2005). Finally, just as there seems to be a bias against exploratory research, similar things can be said for inductive/abductive research. This is echoed in Fidler et al. (2018): "Part of the solution will be to (a) expand what is considered legitimate scientific activity to include exploratory research that is explicitly presented as exploratory and (b) value the inductive and abductive reasoning supporting this work."

This procedure in following confirmatory modeling with exploratory analysis could be conceptualized as following what Cattell (1966) termed the Inductive-Hypothetico-Deductive Spiral (Tellegen and Waller, 2008), which could also be said to follow abductive reasoning (Haig, 2005, 2014). With the abductive theory of method, sets of data are analyzed to detect empirical regularities (robust phenomenon), which are then used to develop explanatory theories to explain their existence. This is followed by constructing plausible models through the use of analogy to relevant domains. Finally, if the explanatory theories become well developed, they are then assessed against rival theories with respect to their explanatory value or goodness (e.g., Ylikoski and Kuorikoski, 2010).

While the majority of psychological research operates from a HD perspective, the new forms of data collection and modeling have motivated increased use of either inductive or abductive reasoning. Particularly with large datasets, it becomes increasingly difficult to have a fully formed theory that can be translated into a data model. Further, viewing more complex algorithms such machine learning from the perspective of hypothetico-deductive perspective can lead to unnecessary (terming machine learning as purely exploratory) and strange (classifying machine learning as EDA along with visual displays of residuals; Fife and Rodgers, 2022) formulations.

Explanation, prediction, description

Similar to how the majority of psychological research has operated from a hypothetico-deductive perspective, thus obviating the necessity of justification, the same could be said for explanatory aims (e.g., Yarkoni and Westfall, 2017). While explanation can be contrasted with description and prediction (Shmueli, 2010; Hamaker et al., 2020; Mõttus et al., 2020), the distinction between description and explanation is often less than clear, and is subject to a researcher's point of view (Wilkinson, 2014; Yarkoni, 2020). While explanation is concerned with understanding underlying mechanisms.[7] this is often seen as intimately linked to theory, as in "we typically need to have a theory about what factors may serve as causes," whereas in descriptive research "we need very little theory to base our research on" (Hamaker et al., 2020, p. 2). Here, we see strong connections to the contrast between exploration and confirmation, as well as between HD and inductive/abductive methods. While there is strong overlap between concepts, describing a study as explanatory clearly orients the reader to the fundamental aim of identifying mechanisms, whereas the concepts of exploration and confirmation are descriptive with respect to theory.

Theory generation, appraisal, and development

The final dimension is denoting whether a study is primarily concerned with theory generation, development, or appraisal (Haig, 2014). Theory appraisal, quite likely the most common stage of research detailed in psychology publications, is traditionally conducted following HD (e.g., see Locke, 2007), outlining the theory in a hypothesis, then followed by a statistical test. This also corresponds almost directly to previous descriptions of confirmation that rely on hypothesis (theory) testing. Theory development could be seen as either confirmatory or exploratory. Confirmatory if hypotheses are concerned with amendments to specific theory, or exploratory if the theory development is based on following Good's (1983) description of EDA as a mechanism to deepen a theory. Finally, exploration aligns almost perfectly with the concept of theory generation, which may be best captured by the previously detailed quote: "Explicit hypotheses tested with confirmatory research usually do not spring from an intellectual void but instead are gained through exploratory research" (p. S64; Jaeger and Halliday, 1998).

Conclusion

Ultimately, the problem with the application of the labels of exploratory/confirmatory can be summarized as an issue in the application of a single dimension solution to multi-dimensional problems. While recent research calls for increased specification on whether the study is exploratory or confirmatory (Wagenmakers et al., 2012; Kimmelman et al., 2014), this will continue to be an overly simple solution. The above sections highlighted deficiencies in their use and how these terms can mask or prevent greater depth in explanation and reporting, particularly in the context of big data. Only the simplest psychological studies could be considered as strictly confirmatory, thus making this exercise futile in generalizing to psychological science broadly.

Each year that goes by results in increased complexity to psychological research, requiring ever more complex levels of decisions made about what variables to collect, which to include in analyses, and what statistical algorithms to use, among many others. Most research cannot and should not be required to have complete theoretical justification for each decision made, as this would severely limit a researcher's level of flexibility and creativity, not to mention foster deceptive research practices and overstated results. In most contexts the terms confirmatory/exploratory simply refer to whether the Hypothetico-Deductive method was followed across the entire study, or specific hypotheses. Criticisms of the HD approach also apply to the use of exploratory/confirmatory, namely that researchers often feel justified in specifying underdeveloped or vague hypotheses and using non-risky tests (i.e., Fidler et al., 2018).

Instead, the rise of more flexible statistical algorithms has been matched by moves away from more traditional Hypothetico-Deductive research. Instead of pigeonholing these new developments in how research conducted into relatively archaic boxes of exploratory or confirmatory research, I advocate for providing detail on how replication/generalizability was addressed statistically, the form of reasoning used in developing the study procedures, whether explanation, prediction, or description is the primary aim, and finally, what stage of theory generation, development or appraisal the research line is in.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary materials, further inquiries can be directed to the corresponding author.

Author contributions

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

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.

[7] This is an overly simple definition. See Wilkinson (2014) for further detail on various types of explanation.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., et al. (2020). A consensus-based transparency checklist. *Nat. Hum. Behav.* 4, 4–6. doi: 10.1038/s41562-019-0772-6

Bailer-Jones, D. M. (2009). *Scientific Models in Philosophy of Science*. Pittsburgh, PA: University of Pittsburgh Press.

Bartha, P. (2013). "Analogy and analogical reasoning," in *The Stanford Encyclopedia of Philosophy*. ed. E. N. Zalta (Stanford, CA: Metaphysics Research Lab)

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods* 2:131.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Borsboom, D., Cramer, A. O., and Kalis, A. (2019). Brain disorders? Not really: why network structures block reductionism in psychopathology research. *Behav. Brain Sci.* 42. doi: 10.1017/s0140525x17002266

Cattell, R. B. (1966). *Handbook of multivariate experimental psychology*. Chicago: Rand McNally.

de Rooij, M., and Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Adv. Methods Pract. Psychol. Sci.* 3, 248–263. doi: 10.1177/2515245919898466

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Depaoli, S., and Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. *Psychol. Methods* 22:240. doi: 10.1037/met0000065

DeYoung, C. G., and Krueger, R. F. (2020). To wish impossible things: on the ontological status of latent variables and the prospects for theory in psychology. *Psychol. Inq.* 31, 289–296. doi: 10.31234/osf.io/4anhr

Etz, A., and Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: psychology. *PLoS One* 11:e0149794. doi: 10.1371/journal.pone.0149794

Fidler, F., Singleton Thorn, F., Barnett, A., Kambouris, S., and Kruger, A. (2018). The epistemic importance of establishing the absence of an effect. *Adv. Methods Pract. Psychol. Sci.* 1, 237–244. doi: 10.1177/2515245918770407

Fife, D. A., and Rodgers, J. L. (2022). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis". *American Psychologist* 77:453.

Franklin, L. R. (2005). Exploratory experiments. *Philos. Sci.* 72, 888–899. doi: 10.1086/508117

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychol. Inq.* 31, 271–288. doi: 10.1080/1047840X.2020.1853461

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., and Dehghani, M. (2018). Dictionaries and distributions: combining expert knowledge and large scale textual data content analysis. *Behav. Res. Methods* 50, 344–361. doi: 10.3758/s13428-017-0875-9

Gelman, A. (2004). Exploratory data analysis for complex models. *J. Comput. Graph. Stat.* 13, 755–779. doi: 10.1198/106186004X11435

Gelman, A., and Loken, E. (2014). The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *Am. Sci.* 102:460. doi: 10.1511/2014.111.460

Goldacre, B. (2009). *Bad Science*. London, England: Fourth Estate.

Good, I. J. (1983). The philosophy of exploratory data analysis. *Philos. Sci.* 50, 283–295. doi: 10.1086/289110

Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* 3, 513–525. doi: 10.1038/s41562-019-0566-x

Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivar. Behav. Res.* 40, 303–329. doi: 10.1207/s15327906mbr4003_2

Haig, B. D. (2014). *Investigating the Psychological World: Scientific Method in the Behavioral Sciences*. Cambridge, MA: MIT Press.

Haig, B. D. (2020). "Big data science: a philosophy of science perspective," in *Big Data in Psychological Research* (Washington, DC: American Psychological Association), 15–33. doi: 10.1037/0000193-002

Hamaker, E. L., Mulder, J. D., and van IJzendoorn, M. H. (2020). Description, prediction and causation: methodological challenges of studying child and adolescent development. *Dev. Cogn. Neurosci.* 46:100867. doi: 10.1016/j.dcn.2020.100867

Harrell, F. (2015). "Regression modeling strategies," *With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. New York: Springer.

Hermida, R. (2015). The problem of allowing correlated errors in structural equation modeling: concerns and considerations. *Comput. Methods Soc. Sci.* 3, 5–17. doi: 10.1037/e518392013-131

Howard, D. (2006). "Lost wanderers in the forest of knowledge: some thoughts on the discovery-justification distinction," in *Revisiting Discovery and Justification: Historical and Philosophical Perspectives on the Context Distinction*. eds. J. Schickore and F. Steinle (Dordrecht: Springer), 3–22.

Jacobson, N. C., Lekkas, D., Huang, R., and Thomas, N. (2021). Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17-18 years. *J. Affect. Disord.* 282, 104–111. doi: 10.1016/j.jad.2020.12.086

Jacobucci, R., Littlefield, A., Millner, A. J., Kleiman, E. M., and Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clin. Psychol. Sci.* 9, 129–134. doi: 10.1177/2167702620954216

Jaeger, R. G., and Halliday, T. R. (1998). On confirmatory versus exploratory research. *Herpetologica* S64–S66.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. New York: springer.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1037/e632032012-001

Kasper, D., and Ünlü, A. (2013). On the relevance of assumptions associated with classical factor analytic approaches. *Front. Psychol.* 4:109. doi: 10.3389/fpsyg.2013.00109

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior* 2, 160–165.

Kimmelman, J., Mogil, J. S., and Dirnagl, U. (2014). Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol.* 12:e1001863. doi: 10.1371/journal.pbio.1001863

Koul, A., Becchio, C., and Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Front. Psychol.* 9:1117. doi: 10.3389/fpsyg.2018.01117

Ledgerwood, A. (2018). The preregistration revolution needs to distinguish between predictions and analyses. *Proc. Natl. Acad. Sci.* 115, E10516–E10517. doi: 10.1073/pnas.1812592115

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., et al. (2019). Robust modeling in cognitive science. *Comput. Brain Behav.* 2, 141–153. doi: 10.1007/s42113-019-00029-y

Lilienfeld, S. O., and Pinto, M. D. (2015). Risky tests of etiological models in psychopathology research: the need for meta-methodology. *Psychol. Inq.* 26, 253–258. doi: 10.1080/1047840X.2015. 1039920

Lindsay, D. S. (2015). Replication in psychological science. *Psychol. Sci.* 26, 1827–1832. doi: 10.1177/0956797615616374

Locke, E. A. (2007). The case for inductive theory building. *J. Manag.* 33, 867–890. doi: 10.1177/0149206307307636

Lovibond, P. F., and Lovibond, S. H. (1995). The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the Beck depression and anxiety inventories. *Behav. Res. Ther.* 33, 335–343. doi: 10.1016/0005-7967(94)00075-U

MacCallum, R. C., Roznowski, M., and Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychol. Bull.* 111:490. doi: 10.1037/0033-2909.111.3.490

Mansell, W., and Huddy, V. (2018). The assessment and modeling of perceptual control: a transformation in research methodology to address the replication crisis. *Rev. Gen. Psychol.* 22, 305–320. doi: 10.1037/gpr0000147

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* 70:487. doi: 10.1037/a0039400

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Curr. Dir. Psychol. Sci.* 5, 11–18. doi: 10.1111/1467-8721.ep10772681

McArdle, J. J. (2012). "Exploratory data mining using CART in the behavioral sciences," in *APA Handbook of Research Methods in Psychology. Data analysis and Research Publication* (Washington, DC: American Psychological Association), 405–421.

McNeish, D., and Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior research methods* 52, 2287–2305.

Meehl, P. E. (1997). "The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions," in *What If There Were No Significance Tests*. eds. L. L. Harlow, S. A. Mulaik and J. H. Steiger (London: Psychology Press)

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv* [Epub ahead of preprint]. doi: 10.48550/arXiv.1301.3781

Moore, D. A. (2016). Pre-register if you want to. *Am. Psychol.* 71, 238–239. doi: 10.1037/a0040195

Morawski, J. (2019). The replication crisis: how might philosophy and theory of psychology be of use? *J. Theor. Philos. Psychol.* 39:218. doi: 10.1037/teo0000129

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., et al. (2020). Descriptive, predictive and explanatory personality research: different goals, different approaches, but a shared need to move beyond the big few traits. *Eur. J. Personal.* 34, 1175–1201. doi: 10.31234/osf.io/hvk5p

Mulkay, M., and Gilbert, G. N. (1981). Putting philosophy to work: Karl Popper's influence on scientific practice. *Philos. Soc. Sci.* 11, 389–407. doi: 10.1177/004839318101100306

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., et al. (2019). Preregistration is hard, and worthwhile. *Trends Cogn. Sci.* 23, 815–818. doi: 10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proc. Natl. Acad. Sci.* 115, 2600–2606.

Nosek, B. A., and Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Obs.* 31. doi: 10.31219/osf.io/2dxu5

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahway: Lawrence Erlbaum Associates.

Ram, N., and Grimm, K. J. (2009). Methods and measures: growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *Int. J. Behav. Dev.* 33, 565–576. doi: 10.1177/0165025409343765

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge.* Chicago, IL: University of Chicago Press.

Rigdon, E. E. (2016). Choosing PLS path modeling as analytical method in European management research: a realist perspective. *Eur. Manag. J.* 34, 598–605. doi: 10.1016/j.emj.2016.05.006

Robinaugh, D., Haslbeck, J., Ryan, O., Fried, E. I., and Waldorp, L. (2020). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspect. Psychol. Sci.* 16, 725–743. doi: 10.1177/1745691620974697

Scheel, A. M., Tiokhin, L., Isager, P. M., and Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspect. Psychol. Sci.* 16, 744–755. doi: 10.1177/1745691620966795

Shmueli, G. (2010). To explain or to predict? *Stat. Sci.* 25, 289–310. doi: 10.2139/ssrn.1351252

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2021). Pre-registration: why and how. *J. Consum. Psychol.* 31, 151–162. doi: 10.1002/jcpy.1208

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Smaldino, P. E. (2017). Models are stupid, and we need more of them. *Comput. Social Psychol.*, 311–331.

Snedecor, G. W., and Cochran, W. G. (1980). *Statistical methods*. IOWA. Iowa State University Press.

Sovacool, B. (2005). Falsification and demarcation in astronomy and cosmology. *Bull. Sci. Technol. Soc.* 25, 53–62. doi: 10.1177/0270467604270151

Suppes, P. (1966). "Models of data," in *Studies in logic and the foundations of mathematics*, vol. *44* (Elsevier), 252–261.

Tellegen, A., and Waller, N. G. (2008). Exploring personality through test construction: Development of the multidimensional personality questionnaire. *The SAGE Handbook of Personality Theory and Assessment* 2, 261–292.

Tukey, John W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley.

van't Veer, A. E., and Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *J. Exp. Soc. Psychol.* 67, 2–12. doi: 10.1016/j.jesp.2016.03.004

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., and Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* 7, 632–638. doi: 10.1177/1745691612463078

Wilkinson, S. (2014). Levels and kinds of explanation: lessons from neuropsychiatry. *Front. Psychol.* 5:373. doi: 10.3389/fpsyg.2014.00373

Yarkoni, T. (2020). Implicit realism impedes Progress in psychology: comment on Fried (2020). *Psychol. Inq.* 31, 326–333. doi: 10.1080/1047840X.2020.1853478

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Ylikoski, P., and Kuorikoski, J. (2010). Dissecting explanatory power. *Philos. Stud.* 148, 201–219. doi: 10.1007/s11098-008-9324-z

Yuan, K. H., Cheng, Y., and Maxwell, S. (2014). Moderation analysis using a two-level regression model. *Psychometrika* 79, 701–732. doi: 10.1007/s11336-013-9357-x

frontiers | Frontiers in Psychology

TYPE Systematic Review
PUBLISHED 12 January 2023
DOI 10.3389/fpsyg.2022.1015921

Check for updates

OPEN ACCESS

EDITED BY
Luca Tummolini,
National Research Council, Italy

REVIEWED BY
Guglielmo Faldetta,
Kore University of Enna, Italy
Jean-Marie C. Chevalier,
Université Paris-Est
Créteil Val de Marne, France

*CORRESPONDENCE
Rehan Ahmad
✉ rahmad05@gmail.com

SPECIALTY SECTION
This article was submitted to
Theoretical and Philosophical Psychology,
a section of the journal
Frontiers in Psychology

RECEIVED 10 August 2022
ACCEPTED 22 December 2022
PUBLISHED 12 January 2023

CITATION
Ahmad R, Nawaz MR, Ishaq MI,
Khan MM and Ashraf HA (2023) Social
exchange theory: Systematic review and
future directions.
Front. Psychol. 13:1015921.
doi: 10.3389/fpsyg.2022.1015921

COPYRIGHT
© 2023 Ahmad, Nawaz, Ishaq, Khan and Ashraf. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Social exchange theory: Systematic review and future directions

Rehan Ahmad[1]*, Muhammad Rafay Nawaz[2], Muhammad Ishtiaq Ishaq[3], Mumtaz Muhammad Khan[1] and Hafiz Ahmad Ashraf[4]

[1]Imperial College of Business Studies, Lahore, Pakistan, [2]Banking and Finance, University of the Punjab, Lahore, Punjab, Pakistan, [3]Quaid-i-Azam University, Islamabad, Pakistan, [4]Management Sciences, University of Central Punjab, Lahore, Punjab, Pakistan

Social exchange theory (SET) is one of the most influential theories in social sciences, which has implications across various fields. Despite its usefulness being a typical social transaction, there is a need to look at it from the lens of psychological transactions to further its evolution and to identify future directions. After generally reviewing 3,649 articles from the Social Science Citation Index and Scopus, a total of 46 articles were selected for final review using a comprehensive systematic review approach. We have highlighted the need for further research in psychological transactions, reciprocity principles, exchange relations, and the impact of various factors on the exchange process. Among other exchange rules (social, economic, and psychological) and transactions (social, economic, and psychological), this research provides an elevation platform for the less explored exchange rules in psychological transactions. Among other theories in the social sciences, social exchange theory is a theory that shadows many other theories under its umbrella.

KEYWORDS

social exchange theory, reciprocity, workplace relations, evolution of social behaviors, social exchange behavior

1. Introduction

Social exchange theory (SET) is one of the gold standards to understand workplace behavior (Cropanzano and Mitchell, 2005). It is such a common phenomenon that is deeply inculcated in our daily lives. Exchanges are not limited to the organizations but extended to our family, friends, and relatives, and that too on a subtle basis. Cropanzano et al. (2017) defined the SET as (i) an initiation by an actor toward the target, (ii) an attitudinal or behavioral response from the target in reciprocity, and (iii) the resulting relationship. Relationships in the corporate world today are becoming increasingly complex (Chernyak-Hai and Rabenu, 2018). Hence, there is a need to update SET with the increasing complexity of how organizations operate and how employees behave (Cooper-Thomas and Morrison, 2019).

Rooted back in the 1920s (Malinowski, 1922; Mauss, 1925), social exchange theory has implications across various fields like social psychology (Homans, 1958; Thibault and Kelley, 1959; Gouldner, 1960), sociology (Blau, 1964), and anthropology (Firth, 1967; Sahlins, 1972). It was Homans (1958), who, for the first time, proposed the idea of "Social behavior as exchange" in the literature, and he further evolved this idea into its elementary forms in 1961. Thibault and Kelley (1959) proposed the converging notion of the "social psychology of groups." Blau (1964) further evolved this idea by presenting the concept of "exchange and power," which refers to the ability of one party to influence another party to do something. Blau highlighted the economic orientation of the theory, while Homans lodged more upon psychological orientation, that is, instrumental behavior. According to a significant contribution by Blau (1964) in literature, social exchange conceived here is limited to actions that are contingent on rewarding reactions from others, and exchange behavior means voluntary actions of individuals that are motivated by the returns they are expected to bring.

Homans (1969) further evolved his study in SET, incorporated sociology, and behavioral psychology concepts and stressed the need for further research on the subject, while Anderson et al. (1969) reinforced the economic implications of the theory. Goode proposed the idea that the role theory and exchange theory were convergent to one another in 1973. Emerson (1976a) suggested that SET is not a theory but a frame covering many theories under its shadow. Other areas analyzed under the light of SET include commitment (Bishop et al., 2000), organizational citizenship behaviors (Organ, 1990), supervisory and organizational support (Ladd and Henry, 2000), and justice (Tepper and Taylor, 2003). Mitchell et al. (2012) proposed the idea of a social life cycle that refers to events/transactions between parties.

Cropanzano et al. (2017) proposed that the action of the first actor is termed initiating action and is divided into positive and negative ones. Positive initiating actions include justice (Cropanzano and Rupp, 2008) and organizational support (Riggle et al., 2009), and negative actions may consist of incivility (Andersson and Pearson, 1999; Pearson et al., 2005), abusive supervision (Tepper et al., 2009), and bullying (Rayner and Keashly, 2005). The resulting response from the target can be classified as behavioral and relational. Subsequently, successful exchanges eventually transform a preliminary economic exchange into a social exchange relationship (Cropanzano et al., 2017). Lyons and Scott (2012) proposed the idea of "homeomorphic reciprocity" which refers to the ability of an employee to receive help or harm shall depend upon the extent to which that employee engages in benefit and harm. Additionally, the behaviors exchanged between an employee and a given coworker should be equivalent, such that engaging in help, but no harm, is associated with receiving support, and engaging in harm, but not help, is associated with receiving harm.

Having such broad applications, according to the study of Cropanzano and Mitchell (2005), the core ideas that comprise SET have yet to be adequately articulated and integrated. Researchers further concluded that SET is a broad framework that can describe almost any finding (Sharpley, 2014; Cropanzano et al., 2017). Such broadness shows the presence of flexibility and variety in SET consequently. At the same time, various researchers embark upon social and economic transactions and exchanges in SET. Based on the call of Cropanzano et al. (2017), this article aims to investigate more upon inactive exchanges, which we termed as psychological exchanges. Active exchanges are visible, while inactive exchanges are less visible and are positive (withholding undesirable behavior) as well as negative (withholding desirable behavior). The shadow nature of the inactive exchanges can turn out to be more damaging for the organization as it is difficult to trace. Moreover, on the basis of the rules of reciprocity, usually more behaviors are inactive and destructive rather than inactive and constructive. Hence, these inactive exchanges are important to explore for a better understanding of SET.

Moreover, building on the definition of SET by Cropanzano et al. (2017), this article further proposes that initiating action, which is found to be explicit, can be implicit, such as a feeling (positive or negative), and can be an outcome of someone's achievement (feeling jealousy at the promotion of a coworker, a psychological exchange). This article comprehensively outlines the evolution of SET and introduces a new dimension in social exchange relationships and ultimately provides future direction for further research.

2. Methods

To understand the social exchange theory and its evolution, one should begin by identifying the roots of the concept and elaborate on the differences and commonalities in the work of various authors in academic literature. The literature highlights different definitions, rules, approaches, and dimensions in the evolution of SET. To understand the concept of SET, three different areas are acknowledged using content analysis of 3,221 articles indexed in the ISI Web of Knowledge and Scopus. The areas are (1) basic concepts of SET as they evolved, (2) exchange rules that govern social exchanges, and (3) evolving dimensions of the exchange relationships. The theoretical framework used in this article is in line with the study of Yadav (2014) and MacInnis (2011), where they propose to differentiate and assimilate particular conceptual goals. We searched the ISI Web of Knowledge and Scopus along with Social Sciences Citation Index from 1920 to 2020 because the concept of SET goes back to 1920.

Search results from the Sciences, Arts, and Humanities Citation Index were eliminated, and the results were filtered for Business and Management, Social Sciences, and Psychology. We used multiple keywords in the ISI search engine in *the topic* field using a complete list of possibilities including "social exchange theory," "exchange relationships," "evolution of social exchange theory," and "exchange relations." These searches returned highly significant empirical and conceptual references ($n = 3{,}221$; Scopus = 1954 and ISI Web of Knowledge = 1,267). After the search, duplicate articles ($n = 1{,}526$) in both databases were deleted.

In the next step, conceptual and empirical articles on SET were separated and analyzed to identify and track evolution patterns, and empirical articles with no theoretical contribution ($n = 1{,}202$) were excluded. In the next phase, those articles were eliminated through contextual analysis that had meager theoretical contributions or available models' allowance ($n = 446$). The purpose of this article was to classify the evolution of SET to propose needed contributions. Hence, after excluding empirical articles and literature reviews with no progression in SET, we ended up with 47 articles (Table 1). Out of the articles that were selected for the final review, two of them were published in the decade between 1920 and 1930, three between 1951 and 1960, five between 1961 and 1970, nine between 1971 and 1980, four between 1981 and 1990, eight between 1991 and 2000, 10 between 2001 and 2010, and nine between 2011 and 2020.

3. Key ideas of set

We shall begin by curating the underlying ideas which comprise SET which involve rules and norms of exchange, resources exchanged, and resulting relationships (Methot et al., 2016; Cropanzano et al., 2017). A comprehensive snapshot of key ideas related to SET across the years is presented in Table 2.

3.1. Rules and norms of exchange

One of the fundamental pillars of SET is that commitment, loyalty, and trust are upshot of evolving relationships with time (Cropanzano and Mitchell, 2005). This pillar demands that parties must show compliance toward specific rules (i.e., rules of exchange). According to Emerson (1976b), such rules form a normative definition of the participants in an exchange relation adopted. Hence, such an exchange principle facilitated avenues for researchers in organizational behavior to further their work (Cropanzano and Mitchell, 2005). Most management research is focused on the potential of reciprocity. Ko and Hur (2014) stressed that other rules of exchange exist that the researchers do not sufficiently explore. This article, therefore, analyzes reciprocity and other less-explored exchange rules.

3.1.1. Reciprocity rules

Gouldner (1960) made a significant contribution to the literature by outlining rules of reciprocity as (a) transaction, (b) belief, and (c) moral norm. The transaction, according to Gouldner (1960), meant interdependent (both dependent on one another) exchanges, and this idea was then reinforced by Molm (1994). A reciprocal exchange due to interdependence curbs risks and supports cooperation, according to Molm (1994), and does not include pronounced bargaining (Molm, 2003). As per the idea, the exchange is a continuous cycle where one party makes a move, and the other reciprocates, and it begins a new cycle of exchanges (Cropanzano and Mitchell, 2005). Suffice it to say that there is a vast literature on the interdependence of exchange and transaction, and reviewing that literature would bypass the scope of this article.

The second rule of reciprocity, that is, reciprocity as belief, revolves around cultural orientation (Gouldner, 1960). This orientation is in line with the idea of karma: You get what you deserve. The idea of a "just world" proposed by Lerner (1980) is consistent with this type of reciprocity. Furthermore, it reduces destructive behavior in people (Bies and Tripp, 1996). Gouldner (1960) speculated that reciprocity is a moral norm and is embedded in humans universally (Tsui and Wang, 2002; Wang et al., 2003). Nevertheless, it is important to note that humans are different, and the way they reciprocate depends heavily on their cultural and individual differences (Parker, 1998; Coyle-Shapiro and Neuman, 2004).

Social psychologists such as Clark and Mills (1979) and Murstein et al. (1977) proposed classifications of individuals based on the degree of reciprocity. They termed the classification "high exchange orientation" (those who readily reciprocate) and "low exchange orientation" (those who do not return or reciprocate less). This unleashed avenues for further research in management as scholars worked on various avenues such as absenteeism (Eisenberger et al., 1986), felt obligation (Eisenberger et al., 2001), citizenship behavior (Witt, 1991), satisfaction and training (Witt and Broach, 1993), performance (Orpen, 1994), union support (Sinclair and Tetrick, 1995), job commitment and satisfaction (Witt et al., 2001), and organizational politics (Andrews et al., 2003).

Many researchers, including Uhl-Bien and Maslyn (2003) and Eisenberger et al. (2004), further classified reciprocity as positive (reciprocating favorable treatment) and negative (reciprocating unfavorable treatment). Cropanzano and Mitchell (2005) called for further investigation into the impact of social exchanges on organizational relationships and also proposed the need for research in unexplored areas such as coworkers, supervisors, and outsiders. Building on previous literature, Cropanzano et al. (2017) proposed that people may not reciprocate the way they wish due to various uncontrollable factors (the presence of inadequate supervision and fewer turnover intentions due to a bad economy). Cropanzano et al. (2017) further added to the literature of SET that reciprocity happens, both explicitly (active exchanges) and implicitly (inactive exchanges). Both forms communicate in exciting ways. For instance, an employee will have high work deviance (implicit) but will not leave the job due to a lousy economy in terms of inactive exchanges (explicit). Moreover, Greco et al. (2019) investigated the reciprocity of negative work behaviors between two parties and reported that negative work behaviors are returned on the similar intensity and capacity between the two parties.

TABLE 1 Evolution of social exchange theory.

Year	Author(s)	Evolution
1920–1930	Malinowski (1922)	The circulating exchange of' valuables in the Archipelagoes of Eastern New Guinea.
	Mauss (1925)	Forms and functions of exchange in Archaic Societies.
1951–1960	Homans (1958)	Social behavior as exchange psychological orientation.
	Thibault and Kelley (1959)	The social psychology of groups.
	Gouldner (1960)	Incorporated types of reciprocity (transaction, belief, moral norm) in the concept of SET.
1961–1970	Blau (1964)	Exchange and power economic orientation.
	Firth (1967)	The implication of SET in anthropology.
	Homans (1969)	Incorporated the concepts of sociology and behavioral psychology.
	Gergen (1969)	Transactions mean interdependent exchanges.
	Anderson et al. (1969)	Reinforced the economic implications of SET.
1971–1980	Meeker (1971)	Proposed six exchange rules as competition, group gain, status consistency, altruism, rationality, and reciprocity.
	Sahlins (1972)	Presented comparison of stone age economics with SET and highlighted implications of SET in anthropology.
	Goode (1973)	Role theory and exchange theory are convergent to one another.
	Emerson (1976a)	SET is not a theory but a frame that covers many theories under its shadow.
	Foa and Foa (1974)	Classifications of exchange resources as status, information, goods, love, money, and services.
	Emerson (1976b)	Exchange relationships are based on the rules of the exchange.
	Clark and Mills (1979)	Classification of individuals based on the degree of reciprocity.
	Foa and Foa (1980)	Classification of exchange resources in two dimensions as economic (tangible) and socioemotional (symbolic).
	Lerner (1980)	The idea of a "just world" in exchange relationships.
1981–1990	Mills and Clark (1982)	Proposed competition and communal exchange relationships.
	Cook et al. (1983)	Concept of terms and rules in social exchange to reach interdependent goals.
	Folger and Konovsky (1989)	SET is beyond the rules of transactions and benefits.
	Organ (1990)	Organizational citizenship behavior in light of SET.
1991–2000	Cropanzano and Baron (1991)	Concept of seeking revenge in an exchange relationship.
	Martin and Harder (1994)	Tangible and symbolic dimensions of exchange resources are based on different exchange rules.
	Molm (1994)	Interdependence in exchanges overcome risks and supports cooperation.
	Chen (1995)	Dimensional classification of exchange resources.
	Batson (1995)	Altruism as an exchange rule.
	Bies and Tripp (1996)	SET concerning justice can reduce destructive behavior in people.
	Bishop et al. (2000)	Organizational commitment in the light of SET.
	Ladd and Henry (2000)	Supervisory and organizational support.
2001–2010	Tsui and Wang (2002)	SET is a moral norm.
	Rhoades and Eisenberger (2002)	Explored exchange relationships as POS and LMX
	Rupp and Cropanzano (2002)	Mediating role of social exchange relationships in predicting workplace outcomes.
	Wang et al. (2003)	SET is embedded in humans universally.
	Molm (2003)	Concept of negotiated exchanges.
	Tepper and Taylor (2003)	Organizational justice in the light of SET.
	Eisenberger et al. (2004)	Concept of positive and negative reciprocity.
	Coyle-Shapiro and Conway (2004)	Employment relationship through the lens of SET.
	Cropanzano and Rupp (2008)	Justice as positive initiating action.

(Continued)

TABLE 1 (Continued)

Year	Author(s)	Evolution
2010–2020	Mitchell et al. (2012)	The social life cycle refers to events/ transactions between parties.
	Lyons and Scott (2012)	Homeomorphic reciprocity.
	Sharpley (2014)	SET as a broad framework that can describe almost any findings
	Ko and Hur (2014)	Rules of exchange introduced in the literature.
	Methot et al. (2016)	The concept of multiplex relations in social exchanges was introduced, which includes formal and informal relations.
	Cropanzano et al. (2017)	Redefined SET as (i) an initiation by an actor toward the target, (ii) an attitudinal or behavioral response from the target in reciprocity, and (iii) the resulting relationship.
	Cropanzano et al. (2017)	Reciprocity happens both explicitly and implicitly. Concept of transactional chains. Addition of activity dimension.
	Cooper-Thomas and Morrison (2019)	Implications of SET in complicated organizational settings.
	Hossen et al. (2020)	Exchange relationships are the results of mutual benefits.

Source: Authors generated this table from searches on the ISI Web of Knowledge and Scopus. All citations in the table are listed in the reference list.

Individual differences in reciprocity are presented in chronological order in Appendix 1.

3.1.2. Negotiated rules and other exchange rules

Parties in a social exchange may negotiate terms or rules to reach interdependent goals (Cook et al., 1983). There is significant literature on the comparison of reciprocal and negotiated exchanges (Molm, 2003). Key findings suggest that better work relations are the outcome of reciprocity than negotiations. Exchange rules other than reciprocity and negotiation gained more attention in literature from sociology and anthropology researchers than from management researchers (Fiske, 1991). One notable study by Meeker (1971) proposed six exchange rules: competition, group gain, status consistency, altruism, rationality, and reciprocity.

According to Meeker (1971), rationality is a thought process asking for justification for various actions taken by a person according to his preferences. Altruism is about being compassionate and kind, where the good of others is essential, even at the cost of ourselves. This sounds uncanny, but the literature supports the take of Meeker (1971) on altruism as an exchange rule (Batson, 1995). Group gain refers to contributions, and everybody takes (benefits) according to their desire. Group gain omits the idea of interpersonal exchanges and extends the horizon toward group exchanges. Status consistency is also called rank equilibrium, where the disunion of benefits depends upon one's standing in a social group. Lind (1995) experimented with and supported this exchange rule.

Competition is directly the opposite of altruism, where altruism is about benevolence, and competition is about self-seeking behavior (Meeker, 1971). This opened doors for research on modern-day variables in organizational behavior such as workplace envy (Ahmad et al., 2020), organizational politics, and political skills. The study of Meeker (1971) also strengthened the idea of seeking revenge in an exchange relationship (Cropanzano and Baron, 1991; Turillo et al., 2002). A great deal of literature exists on reciprocity as a rule of exchange. Still, there are other rules, such as group gain, status consistency, competition, altruism, and rationality, which require attention and investigation. Exploring these will open doors to fathom the process of social exchanges, which is still unexplored to a great deal (Cropanzano and Mitchell, 2005). Moreover, there is a possibility that multiple exchange rules are employed at once.

3.2. The resources of exchange

Foa and Foa (1974) proposed classifications of exchange resources as status, information, goods, love, money, and services. These resources can be termed as benefits that a person seeks in social exchange and can be further classified into two dimensions economic (tangible) and socioemotional resources (symbolic) (Foa and Foa, 1980). Both dimensions work on different exchange rules (Martin and Harder, 1994). Resources and their dimensional classification are still not sufficiently explored and are open for further investigation. Furthermore, the relationship between types of resources and the type of relationship is also an open area for research (Cropanzano and Mitchell, 2005).

3.3. Resulting relationships: Social exchange relationships

Workplace relationships are the most explored area in management research (Coyle-Shapiro and Conway, 2004). However, much of the research on exchange relations is done in employer–employee relations (Blau, 1964). His study is based on the premise that much of social relations are based on unspecified obligations. This makes the relations more casual while successful

TABLE 2 Key ideas related to SET.

Key ideas	Authors
Rules and norms of exchange	Cropanzano and Mitchell (2005)
	Emerson (1976b)
	Ko and Hur (2014)
Reciprocity rules	Gouldner (1960)
	Molm (1994)
	Molm (2003)
	Lerner (1980)
	Bies and Tripp (1996)
	Cropanzano et al. (2017)
Rules of exchange	Cook et al. (1983)
	Molm (2003)
	Meeker (1971)
	Batson (1995)
Resources of exchange	Foa and Foa (1974)
	Foa and Foa (1980)
	Cropanzano and Mitchell (2005)
Social exchange relationships	Shore and Coyle-Shapiro (2003)
	Blau (1964)
	Mills and Clark (1982)
	Cropanzano and Mitchell (2005)
	Eisenberger et al. (2004)
	Molm (2003)
	Eisenberger et al. (2004)
	Methot et al. (2016)
	Cooper-Thomas and Morrison (2019)
	Cropanzano et al. (2017)
Transactions and exchange relationships	Cropanzano and Mitchell (2005)
	Cropanzano et al. (2017)
	Cropanzano et al. (2017)

All citations in the table are listed in the reference list.

exchanges are based on the commitment between parties. Blau (1964) also considered relations as transactions. Mills and Clark (1982) further contributed to the literature by proposing two types of exchange relationships. One is exchange relations based on competition, and the others are communal relations based on benevolence. Organ (1990) found that SET is beyond the rules of transactions and benefits, and this extended the scope for further research in SET.

Suffice it to note that relations are termed as associations between partners, which can be institutions and individuals (Cropanzano and Mitchell, 2005). Although much of the research is done on exploring the relations between institutions and individuals such as employing organizations (Moorman et al., 1998), customers (Houston et al., 1992), and suppliers (Perrone et al., 2003), the literature is comparatively silent on the area of individual relationships in an organizational setting such as peer relations. Notable work in management is done in terms of exchange relationships which are perceived organizational support (POS), Leader–Member Exchange (LMX; Eisenberger et al., 2004), support to commitment (Eisenberger et al., 1990), team support and organizational support (Bishop et al., 2000), supervisor support (Masterson et al., 2000), and trust (Dirks and Ferrin, 2002).

It is also important to state that relationships develop over time ranging from premature relations (Molm, 2003) to mature ones (Eisenberger et al., 2004). Building on the premise of increasingly complex relationships at the workplace, Methot et al. (2016) introduced the term "multiplex" relations at the workplace, which include both formal (work-related) and informal (friendship) elements. Such relations cover both positive (e.g., emotional support) and negative (e.g., emotional exhaustion) aspects. Cooper-Thomas and Morrison (2019) identified that it is not clear how SET might apply in conditions where positive and negative exchanges are simultaneously taking place.

As multiple behaviors are exchanged in the workplace, Cropanzano et al. (2017) tossed the term "transactional chains" through which relationships are developed over time through various exchanges. If we want to understand the form of a relationship, we must understand the principal transaction of resources responsible for a particular relationship. Building on the need to understand SET in further detail highlighted by Cropanzano et al. (2017) and Cooper-Thomas and Morrison (2019), we shall elaborate on the transactions and resulting exchange relationships.

3.3.1. Transactions and exchange relationships

Cropanzano and Mitchell (2005) highlighted two distinguishing aspects of relationships in the literature. One aspect is a relationship as the series of interdependent transactions transpires to interpersonal attachment, which is a relationship. Alternatively, another element is the interpersonal relationship that originates from interdependent exchanges. It is essential to distinguish the relationship from the transaction process because of its interchangeability. The nature of the relationship between two parties is dictated by the process of exchange or the benefits they exchange between them. When a series of exchanges happen, it becomes rather challenging to find which exchange caused the relationship.

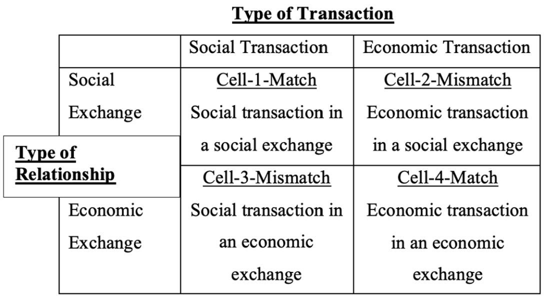
Researchers separated the form of exchange from the exchange relationship presented in Figure 1. Cells 1 and 4 can be termed *matches* as the form of transaction coinciding with the relationship. The situation in Cell 2, where the social exchange relationship coincides with the economic transaction, could reap both risks and rewards. For instance, social relations are at greater risk in economic exchanges, and hence, economic exchanges can pose a more significant threat to relationships (clashes in the inheritance among family members). Alternatively, while considering rewards, greater trust and stronger relationships can be an outcome for such exchanges (father giving money to son and not asking for details). Cell 3 presents the unusual case of emotional labor where employees from the hospitality industry or health workers attend to the emotional needs of their clients or patients for money (economic transaction).

People working in mental asylums display such behaviors to fulfill their professional duties. Similarly, people working in the hotel and hospitality sector are expected to be friendly with their clients. It is tricky and stressful to share such emotions with others, expected to be family members or other loved ones. While keeping in view, the vagueness of the concept of relationships in SET, Cropanzano and Mitchell (2005) highlighted two distinct conceptual dimensions of the relationship. One is a sequence of inter-related exchanges, and the other is relationships as an outcome of codependent exchanges. These are termed transactional and interpersonal relationships in the literature. When relationships seem to transcend over one another, it becomes more challenging to define them. It is essential to understand that two different things can be exchanged through various means among two different parties.

4. Discussion: Beyond socio-economic transactions

Building on the aforementioned model, we propose that while looking beyond the lens of social and economic transactions and exchanges, relationships are also psychological. This premise is based on the idea of implicit or inactive exchanges proposed by Cropanzano et al. (2017). The concept of psychological capital (Luthans et al., 2007) also supports this idea, and exchanges in such relations can be termed psychological exchanges. Referring to Figure 2, Cells 1, 2, 4, and 5 are similar to Cells 1, 2, 3, and 4 in Figure 1. Unique cells in Figure 2 are Cells 3, 6, 7, 8, and 9. Cell 9 is a matching cell coinciding psychological transaction with a psychological exchange relationship. Let us first hone ourselves with the idea of psychological transactions.

To start with, psychological transactions are usually inactive exchanges. From this dimension, it sounds easier to draw that psychological exchange relations are inactive relations, which is incorrect. Psychological relations are based on the *understanding* between the two parties. From "understanding," it means how well parties in a social exchange know each other. This, according to the empirical evidence, indicates that parties develop relationships after being involved in a series of exchanges, and eventually, they develop a relationship so good that they can understand each other on psychological fronts as well. Nevertheless, this is not true as Cell 3 clarifies that psychological transactions may not necessarily occur in every social relationship.



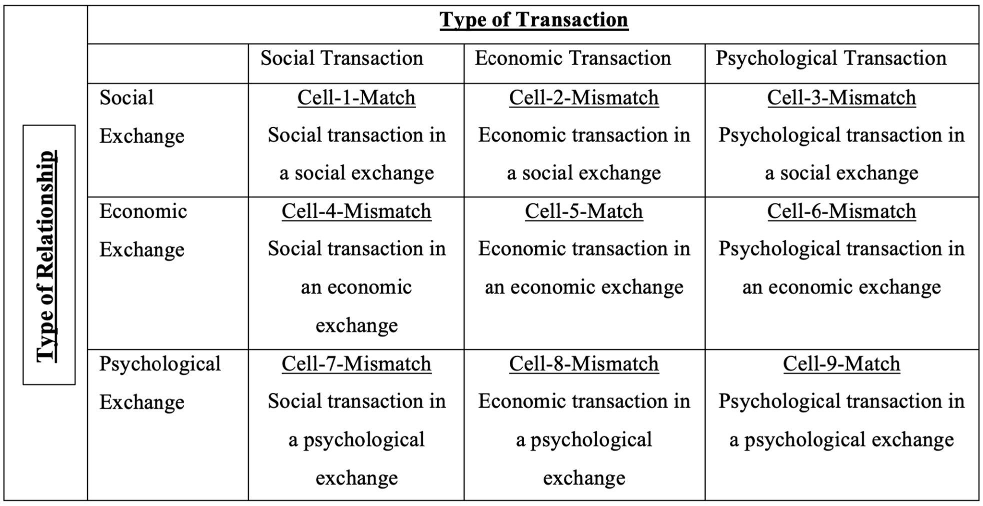
		Type of Transaction	
		Social Transaction	Economic Transaction
Type of Relationship	Social Exchange	Cell-1-Match Social transaction in a social exchange	Cell-2-Mismatch Economic transaction in a social exchange
	Economic Exchange	Cell-3-Mismatch Social transaction in an economic exchange	Cell-4-Match Economic transaction in an economic exchange

FIGURE 1
Relationships of transactions in exchanges.

Putting it further, it is challenging to find like-minded people with whom our mental chemistry aligns. Referring to Cell 6, which draws a dimension about the psychological transaction in an economic relationship, it is evident that psychological transactions do occur during economic relations, but such transactions are usually dubious. The reason for this is that such transactions are generally solitary and not dyad. Due to this attribute, past researchers called them inactive exchanges. Cell 7 presents the case of clinical psychology, where psychiatrists develop a psychological relationship with patients or subjects in a social setting.

Similarly, researchers also fall into this category to build empathy through social transactions to collect data. Cell 8 is similar to Cell 7, and diffusion can be drawn in the *intent*. Cell 7 refers to social welfare, while Cell 8 refers to economic return. If a researcher is working on a social problem or aiming to find a cure for a disease such as COVID-19 without aiming for lucrative gains, he will fall into Cell 7. On the contrary, if Toyota launches an electric vehicle or Philips launches a light bulb that consumes less electricity with a pure aim to sell these products to those consumers who want to save on their gas or electricity bills, they would fall in the Cell 8. If a transaction is taken as a relationship, then successful exchanges will be accepted as its outcome. It works both ways, from transactions in relations to relations in transactions (Figure 2).

To explain how psychological transaction and psychological exchange relations work, the model by Foa and Foa (1980) comes to rescue from the literature. This model aligns a variety of resources according to different relationships, such as causal and universal. Causal relations complement universal resources, while intimate relations complement particularistic resources. Interestingly, a universal benefit paves the way for particularistic use, and this is how relationships become an outcome of reciprocal exchanges. Hence to understand this concept of exchange, we need to further our understanding related to exchange models. As to further contribution to SET literature, two models are proposed below to provide conceptual support to the dimensions of psychological transactions and psychological exchange relationships.



		Type of Transaction		
		Social Transaction	Economic Transaction	Psychological Transaction
Type of Relationship	Social Exchange	Cell-1-Match Social transaction in a social exchange	Cell-2-Mismatch Economic transaction in a social exchange	Cell-3-Mismatch Psychological transaction in a social exchange
	Economic Exchange	Cell-4-Mismatch Social transaction in an economic exchange	Cell-5-Match Economic transaction in an economic exchange	Cell-6-Mismatch Psychological transaction in an economic exchange
	Psychological Exchange	Cell-7-Mismatch Social transaction in a psychological exchange	Cell-8-Mismatch Economic transaction in a psychological exchange	Cell-9-Match Psychological transaction in a psychological exchange

FIGURE 2
Proposed model of transactions and exchanges.

4.1. Nature of relations affects the psychological exchanges

Eisenberger et al. (2001) suggested that employees in an organization can exchange commitment in the reciprocation of organizational support. This finding allowed us to build our argument that the nature of relations between parties who participate in an exchange process can affect psychological exchanges. In other words, the closer the relationship between the two parties (pluralistic exchanges), the more there will be psychological exchanges. The key term to note here is "close," which means seeing someone like peers or classmates every day. Furthermore, the achievement of a friend or classmate who went abroad will affect us less than someone we see every day.

This happens because of the social comparison we do with people near us. Hence, social distance or space between the parties does affect the relationship between them. Moreover, such a relationship will directly impact the intensity or type of psychological exchanges between them. It is important to note that not only the positive relationship enables the possibility of psychological exchanges, but it can also have a similar impact in terms of hostile relations as well. Similarly, a positive relationship does not necessarily mean that there will be only complementary psychological exchanges; negative psychological exchanges can also occur. For instance, you are feeling jealous about the good grades of your best friend. But such a psychological exchange would be different from the one you would have against someone in the class you dislike.

4.2. Psychological exchanges affect the nature of relations

Psychological exchanges in an organization are not a one-time thing but a continuous process like climbing a ladder. In other words, it constitutes a series of transactions between parties in a work setting. Hence, the output of a transaction today will form the psychological resource (both positive and negative) that can be exchanged tomorrow or anytime in the future. Therefore, psychological exchanges can form the basis of relationships between the parties. Positive psychological exchanges become a reason for positive relations, and negative psychological exchanges can cause negative associations (rivalry—usually between coworkers).

It is imperative to note that the exchange timing plays a significant role in forming the relations between parties. This timing of exchange dimension is coherent with the model of LMX development proposed by Uhl-Bien and Maslyn (2003). This model suggests that leaders and members start their relationship journey by testing one another in terms of obligations, and the quality of relations depends upon the reciprocity of commitments. Suffice it to say that positive psychological exchanges result in the exchange of positive psychological resources. Similarly, negative psychological exchanges result in the exchange of harmful psychological resources, which impact resulting relationships.

5. Recommendations and future directions

Having its roots in the 1920s (Malinowski, 1922; Mauss, 1925), the scope and foundations of SET are yet to be sufficiently explored. Management researchers have characteristics of a variety and multiple applications and are doing injustice with this theory in two ways. First, they lack the indulgent understanding of ideas that set the foundations of SET. Second, limited avenues are being explored in the research as reciprocity principles and economic orientation of SET. Cropanzano et al. (2017) investigated that people may not reciprocate according to their wishes due to certain uncontrollable factors. Cooper-Thomas and Morrison (2019) identified that it is not clear how SET might apply in conditions where positive and negative exchanges are simultaneously taking place.

We believe that this article shall help address both shortfalls as it adopts a meek way to outline the evolution of SET and identify essential areas where researchers can direct their future efforts. This article shall help dramatically evolve the theory by revising existing concepts, orientations, and forming new ones. According to Eisenberger et al. (1986) and Graen and Scandura (1987), SET comprises two types of social exchanges. First is perceived organizational support (POS) that emphasizes employee–organization exchange relationships.

The second is the exchange between the leader and member, which elaborates on the interaction between the supervisor and the employee through the exchange of resources (Lee and Duffy, 2019). In both types of exchanges, resulting relationships work as a cynosure of the exchange process. Consequently, the understanding of SET would remain meager if we could not hone the idea of exchanges and resulting relationships. This article pronounced the social and economic transactions and exchanges from the literature and proposed a new *psychological* dimension with empirical and conceptual justifications. This idea is similar to Cropanzano et al. (2017), who introduced the concept of active and inactive exchanges, which revolutionized the whole notion of SET.

According to these dimensions, exchanges in organizational settings happen both explicitly (active exchanges) and implicitly (inactive exchanges). More notably, in the presence of uncontrollable factors, employees will still reciprocate but implicitly. The idea of how employees may get involved in inactive exchanges, even in the absence of uncontrollable factors, is another open avenue for future research. Take an instance of workplace envy: Workplace envy is an inactive exchange (beneficial or costly) of an employee in an organizational setting. It is a feeling that could be visible through active exchanges.

Building on these developments, this study proposes that social exchange may not necessarily be dyadic; it can be individualistic or monotonous where an employee feels on his own. The role of psychological transactions and resulting psychological exchange relationships can be understood from a case as simple as an employee feeling jealous about the achievement of a coworker. This dimension is inevitable, and it nulls the first part of the definition of SET, that is, *initiation by an actor*. This is because no one is initiating, and an employee envies himself or inactive exchange is taking place. Future studies should help to unveil this process of SET in further detail. Moreover, the current study focused on organizational exchanges and resulting relationships, and future research efforts can be directed toward social exchanges among family, friends, and relatives to improve the understanding and scope of SET.

It is also pertinent to note that negative emotions and feelings may be controlled through specific skills such as political skills and social skills. While there is much research on social exchanges in organizational relationships, areas of coworkers, supervisors, and outsiders are yet to be sufficiently explored. Moreover, Foa and Foa (1980) proposed classifications of exchange resources as status, information, goods, love, money, and services. These resources can be further classified into two dimensions economic (tangible) and socioemotional resources (symbolic). On account of social exchange relationships, much of the research is done on exploring relations among institutions and individuals (Moorman et al., 1998), customers (Houston et al., 1992), and suppliers (Perrone et al., 2003), whereas literature is comparatively silent on the area of interpersonal relationships in an organizational setting.

There are exchange rules beyond reciprocity, exchange resources above money, and trust, and there are types of relationships other than social, economic, and psychological that need to be explored. These resources and their impact on social relationships are also unexplored areas asking for attention from the researchers. In addition to the above discussion, the following points can pave the way for a better understanding of SET and future research.

(a) It is unnecessary for a social exchange process that a positive initiating action would generate a positive response.
(b) Positive initiating action may not form a positive relationship.
(c) Positive initiating action may not always form a positive relationship, and it can be negative too.
(d) With changing workplace landscape, relationships are becoming increasingly complex in modern organizations; hence, relations are increasingly affecting the modern exchange process.
(e) An implicit initiating action can cause implicit and explicit behavioral responses.
(f) In some social instances, such as envy, the exchange process can be hidden, and hence, an actual exchange process could be altered with a fabricated exchange process.

6. Conclusion

While SET is evolving, it is inviting researchers to explore various related avenues. Thus, a broad theory that can shadow many other theories under its umbrella can describe multiple social phenomena. This article provided comprehensive commentary about how SET evolved and recent progressions, and it also provides fruit of thought on the psychological dimension that exists under the disguise of inactive exchanges. Beyond social and economic transactions, the idea and implications of psychological transactions are proposed in this article. Based on the idea of inactive exchanges, it is also proposed that other than reciprocity, other less explored exchange rules are dominant in psychological transactions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RA and MN: concept development and systematic review strategy and final write-up. MI and MK: downloading and reviewing manuscript to be selected for the final review. HA: language of the manuscript, bibliography, and final formatting and review. All authors contributed to the article and approved the submitted version.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmad, R., Khan, M. M., and Ishaq, M. I. (2020). The role of envy and psychological capital on performance in the banking industry of Pakistan. *Pak. Soc. Sci. Rev.* 4, 96–112. doi: 10.35484/pssr.2020(4-IV)07

Anderson, B., Berger, J., Zelditch, M.Jr., and Cohen, B. P. (1969). Reactions to inequity. *Acta Sociol.* 12, 1–12. doi: 10.1177/000169936901200101

Andersson, L. M., and Pearson, C. M. (1999). Tit for tat? The spiraling effect of incivility in the workplace. *Acad. Manag. Rev.* 24, 452–471. doi: 10.5465/amr.1999.2202131

Andrews, M. C., Witt, L. A., and Kacmar, K. M. (2003). The interactive effects of organizational politics and exchange ideology on manager ratings of retention. *J. Vocat. Behav.* 62, 357–369. doi: 10.1016/S0001-8791(02)00014-3

Batson, C. D. (1995). "Prosocial motivation: why do we help others?" in *Advanced Social Psychology*. ed. A. T. Tesser (New York: McGraw-Hill), 332–381.

Bies, R. J., and Tripp, T. M. (1996). "Beyond distrust: getting even and the need for revenge" in *Trust in Organizations*. eds. R. M. Kramer and T. Tyler (Thousand Oaks, CA: Sage), 246–260.

Bishop, J. W., Scott, K. D., and Burroughs, S. M. (2000). Support, commitment, and employee outcomes in a team environment. *J. Manag.* 26, 1113–1132. doi: 10.1177/014920630002600603

Blau, P. M. (1964). *Exchange and Power in Social Life*. New York: John Wiley.

Chen, C. C. (1995). New trends in rewards allocation preferences: a Sino-U.S. comparison. *Acad. Manag. J.* 38, 408–428. doi: 10.5465/256686

Chernyak-Hai, L., and Rabenu, E. (2018). The new era workplace relationships: is social exchange theory still relevant? *Ind. Organ. Psychol.* 11, 456–481. doi: 10.1017/iop.2018.5

Clark, M. S., and Mills, J. (1979). Interpersonal attraction in exchange and communal relationships. *J. Pers. Soc. Psychol.* 37, 12–24. doi: 10.1037/0022-3514.37.1.12

Cook, K. S., Emerson, R. M., and Gillmore, M. R. (1983). The distribution of power in exchange networks: theory and experimental results. *Am. J. Sociol.* 89, 275–305. doi: 10.1086/227866

Cooper-Thomas, H. D., and Morrison, R. L. (2019). Give and take: needed updates to social exchange theory. *Ind. Organ. Psychol.* 11, 493–498. doi: 10.1017/iop.2018.101

Cotterell, N., Eisenberger, R., and Speicher, H. (1992). Inhibiting effects of reciprocation wariness on interpersonal relationships. *J. Pers. Soc. Psychol.* 62, 658–668. doi: 10.1037/0022-3514.62.4.658

Coyle-Shapiro, J. A.-M., and Conway, N. (2004). "The employment relationship through the lens of social exchange theory" in *The Employment Relationship: Examining Psychological and Contextual Perspectives*. eds. J. Coyle-Shapiro, L. M. Shore, M. S. Taylor and L. Tetrick (Oxford: Oxford University Press), 5–28.

Coyle-Shapiro, J. A.-M., and Neuman, J. H. (2004). The psychological contract and individual differences: the role of exchange and creditor ideologies. *J. Vocat. Behav.* 64, 150–164. doi: 10.1016/S0001-8791(03)00031-9

Cropanzano, R., Anthony, E. L., Daniels, S. R., and Hall, A. V. (2017). Social exchange theory: a critical review with theoretical remedies. *Acad. Manag. Ann.* 11, 479–516. doi: 10.5465/annals.2015.0099

Cropanzano, R., and Baron, R. A. (1991). Injustice and organizational conflict: the moderating effect of power restoration. *Int. J. Confl. Manag.* 2, 5–26. doi: 10.1108/eb022691

Cropanzano, R., and Mitchell, M. S. (2005). Social exchange theory: an interdisciplinary review. *J. Manag.* 31, 874–900. doi: 10.1177/0149206305279602

Cropanzano, R., and Rupp, D. E. (2008). "Social exchange theory and organizational justice: job performance, citizenship behaviors, multiple foci, and a historical integration of two kinds of literature" in *Research in Social Issues in Management: Justice, Morality, and Social Responsibility*. eds. S. W. Gilliland, D. P. Skarlicki and D. D. Steiner (Greenwich, CT: Information Age Publishing)

de Ruyter, K., and Wetzels, M. (2000). Determinants of a relational exchange orientation in the marketing-manufacturing interface. *J. Manag. Stud.* 37, 257–276. doi: 10.1111/1467-6486.00180

Dirks, K. T., and Ferrin, D. L. (2002). Trust in leadership: meta-analytic findings and implications for research and practice. *J. Appl. Psychol.* 87, 611–628. doi: 10.1037/0021-9010.87.4.611

Eisenberger, R., Armeli, S., Rexwinkel, B., Lynch, P. D., and Rhoades, L. (2001). Reciprocation of perceived organizational support. *J. Appl. Psychol.* 86, 42–51. doi: 10.1037/0021-9010.86.1.42

Eisenberger, R., Cotterell, N., and Marvel, J. (1987). Reciprocation ideology. *J. Pers. Soc. Psychol.* 53, 743–750. doi: 10.1037/0022-3514.53.4.743

Eisenberger, R., Fasolo, P., and Davis-LaMastro, V. (1990). Perceived organizational support and employee diligence, commitment, and innovation. *J. Appl. Psychol.* 75, 51–59. doi: 10.1037/0021-9010.75.1.51

Eisenberger, R., Huntington, R., Hutchison, S., and Sowa, D. (1986). Perceived organizational support. *J. Appl. Psychol.* 71, 500–507. doi: 10.1037/0021-9010.71.3.500

Eisenberger, R., Lynch, P., Aselage, J., and Rohdieck, S. (2004). Who takes the most revenge? Individual differences in negative reciprocity norm endorsement. *Pers. Soc. Psychol. Bull.* 30, 787–799. doi: 10.1177/0146167204264047

Emerson, R. M. (1976a). Social exchange theory. *Annu. Rev.* 2, 335–362.

Emerson, R. M. (1976b). Imperial administration as an exchange network; the length of dynastic rule in the Mugha1 empire. *Inst. Social. Res. Univ. Wash.*

Firth, R. (1967). *Themes in Economic Anthropology*. London: Tavistock.

Fiske, A. P. (1991). *Structures of Social Life: The Four Elementary Forms of Human Relations*. New York: Free Press.

Foa, U. G., and Foa, E. B. (1974). *Societal Structures of the Mind*. Springfield, IL: Charles C Thomas.

Foa, U. G., and Foa, E. B. (1980). "Resource theory: interpersonal behavior as exchange" in *Social Exchange*. eds. K. J. Gergen, M. S. Greenberg and R. H. Willis (Boston, MA: Springer)

Folger, R., and Konovsky, M. A. (1989). Effects of procedural and distributive justice on reactions to pay raise decisions. *Acad. Manag. J.* 32, 115–130. doi: 10.5465/256422

Gallucci, M., and Perugini, M. (2003). Information seeking and reciprocity: a transformational analysis. *Eur. J. Soc. Psychol.* 33, 473–495. doi: 10.1002/ejsp.156

Gergen, K. J. (1969). *The Psychology of Behavioral Exchange*. Reading, MA: Addison-Wesley.

Goode, W. J. (1973). *Explorations in Social Theory* New York: Oxford Theory University Press.

Gouldner, A. W. (1960). The norm of reciprocity: a preliminary statement. *Am. Sociol. Rev.* 25, 161–178. doi: 10.2307/2092623

Graen, G. B., and Scandura, T. A. (1987). Toward a psychology of dyadic organizing. *Res. Organ. Behav.* 9, 175–208.

Greco, L. M., Whitson, J. A., O'Boyle, E. H., Wang, C. S., and Kim, J. (2019). An eye for an eye? A meta-analysis of negative reciprocity in organizations. *J. Appl. Psychol.* 104, 1117–1143. doi: 10.1037/apl0000396

Homans, G. C. (1958). Social behavior as exchange. *Am. J. Sociol.* 63, 597–606. doi: 10.1086/222355

Homans, G. C. (1969). "The sociological relevance of behaviourism" in *Behavioural Sociology: The Experimental Analysis of Social Process*. eds. R. L. Burgess and BushellD. Jr. (New York: Columbia University Press)

Hossen, M. M., Chan, T. J., and Mohd Hasan, N. A. (2020). Mediating role of job satisfaction on internal corporate social responsibility practices and employee engagement in higher education sector. *Contemp. Manag. Res.* 16, 207–227. doi: 10.7903/cmr.20334

Houston, F. S., Gassenheimer, J. B., and Maskulka, J. M. (1992). *Marketing Exchange Transactions and Relationships*. Westport, CT: Quorum Books.

Ko, J., and Hur, S. (2014). The impacts of employee benefits, procedural justice, and managerial trustworthiness on work attitudes: integrated understanding based on social exchange theory. *Public Adm. Rev.* 74, 176–187. doi: 10.1111/puar.12160

Ladd, D., and Henry, R. A. (2000). Helping coworkers and helping the organization: the role of support perceptions, exchange ideology, and conscientiousness 1. *J. Appl. Soc. Psychol.* 30, 2028–2049. doi: 10.1111/j.1559-1816.2000.tb02422.x

Lee, K., and Duffy, M. K. (2019). A functional model of workplace envy and job performance: when do employees capitalize on envy by learning from envied targets? *Acad. Manag. J.* 62, 1085–1110. doi: 10.5465/amj.2016.1202

Lerner, M. J. (1980). "The belief in a just world" in *The Belief in a Just World: A Fundamental Delusion*. ed. M. J. Lerner (Boston, MA: Springer), 9–30.

Lind, E. A. (1995). "Justice and authority relations in organizations" in *Organizational Politics, Justice, and Support: Managing the Social Climate of the Workplace*. eds. R. Cropanzano and M. K. Kacmar (Westport, CT: Quorum Books), 83–96.

Luthans, F., Avolio, B. J., Avey, J. B., and Norman, S. M. (2007). Positive psychological capital: measurement and relationship with performance and satisfaction. *Pers. Psychol.* 60, 541–572. doi: 10.1111/j.1744-6570.2007.00083.x

Lynch, P. D., Eisenberger, R., and Armeli, S. (1999). Perceived organizational support: inferior versus superior performance by wary employees. *J. Appl. Psychol.* 84, 467–483. doi: 10.1037/0021-9010.84.4.467

Lyons, B. J., and Scott, B. A. (2012). Integrating social exchange and affective explanations for the receipt of help and harm: a social network approach. *Organ. Behav. Hum. Decis. Process.* 117, 66–79. doi: 10.1016/j.obhdp.2011.10.002

MacInnis, D. J. (2011). A framework for conceptual contributions in marketing. *J. Mark.* 75, 136–154. doi: 10.1016/j.obhdp.2011.10.002

Malinowski, B. (1922). *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. London: Routledge

Martin, J., and Harder, J. W. (1994). Bread and roses: justice and the distribution of financial and socioemotional rewards in organizations. *Soc. Justice Res.* 7, 241–264. doi: 10.1007/BF02334833

Masterson, S. S., Lewis, K., Goldman, B. M., and Taylor, M. S. (2000). Integrating justice and social exchange: the differing effects of fair procedures and treatment on work relationships. *Acad. Manag. J.* 43, 738–748. doi: 10.5465/1556364

Mauss, M. (1925). *The Gift: Forms and Functions of Exchange in Archaic Societies*. New York: The Norton Library.

Meeker, B. F. (1971). Decisions and exchange. *Am. Sociol. Rev.* 36, 485–495. doi: 10.2307/2093088

Methot, J. R., Lepine, J. A., Podsakoff, N. P., and Christian, J. S. (2016). Are workplace friendships a mixed blessing? Exploring tradeoffs of multiplex relationships and their associations with job performance. *Pers. Psychol.* 69, 311–355. doi: 10.1111/peps.12109

Mills, J., and Clark, M. S. (1982). Exchange and communal relationships. *Rev. Pers. Soc. Psychol.* 3, 121–144.

Mitchell, M. S., Cropanzano, R. S., and Quisenberry, D. M. (2012). "Social exchange theory, exchange resources, and interpersonal relationships: a modest resolution of theoretical difficulties" in *Handbook of Social Resource Theory*. eds. K. Törnblom and A. Kazemi (New York, NY: Springer), 99–118.

Molm, L. D. (1994). Dependence and risk: transforming the structure of social exchange. *Soc. Psychol. Q.* 57, 163–176. doi: 10.2307/2786874

Molm, L. D. (2003). Theoretical comparisons of forms of exchange. *Sociol Theory* 21, 1–17. doi: 10.1111/1467-9558.00171

Moorman, R. H., Blakely, G. L., and Niehoff, B. P. (1998). Does perceived organizational support mediate the relationship between procedural justice and organizational citizenship behavior? *Acad. Manag. J.* 41, 351–357. doi: 10.5465/256913

Murstein, B. I., Cerreto, M., and Mac Donald, M. G. (1977). A theory and investigation of the effect of exchange orientation on marriage and friendship. *J. Marriage Fam.* 39, 543–548. doi: 10.2307/350908

Organ, D. W. (1990). The motivational basis of organizational citizenship behavior. *Res. Organ. Behav.* 12, 43–72.

Orpen, C. (1994). The effects of exchange ideology on the relationship between perceived organizational support and job performance. *J. Soc. Psychol.* 134, 407–408. doi: 10.1080/00224545.1994.9711749

Parker, B. (1998). *Globalization: Managing Across Boundaries*. London: Sage.

Pearson, C. M., Andersson, L. M., and Porath, C. L. (2005). "Workplace incivility" in *Counterproductive Work Behavior: Investigations of Actors and Targets*. eds. S. Fox and P. E. Spector (Washington, DC: American Psychological Association), 177–200.

Perrone, V., Zaheer, A., and McEvily, B. (2003). Free to be trusted? Organizational constraints on trust in boundary spanners. *Organ. Sci.* 14, 422–439. doi: 10.1287/orsc.14.4.422.17487

Perugini, M., and Gallucci, M. (2001). Individual differences and social norms: the distinction between reciprocators and prosocial. *Eur. J. Personal.* 15, S19–S35. doi: 10.1002/per.419

Rayner, C., and Keashly, L. (2005). "Bullying at work: a perspective from Britain and North America" in *Counterproductive work behavior: Investigations of actors and targets*. eds. S. Fox and P. E. Spector (Washington, DC: American Psychological Association), 271–296.

Rhoades, L., and Eisenberger, R. (2002). Perceived organizational support: a review of the literature. *J. Appl. Psychol.* 87, 698–714. doi: 10.1037//0021-9010.87.4.698

Riggle, R. J., Edmondson, D. R., and Hansen, J. D. (2009). A meta-analysis of the relationship between perceived organizational support and job outcomes: 20 years of research. *J. Bus. Res.* 62, 1027–1030. doi: 10.1016/j.jbusres.2008.05.003

Rupp, D. E., and Cropanzano, R. (2002). The mediating effects of social exchange relationships in predicting workplace outcomes from multifocal organizational justice. *Organ. Behav. Hum. Decis. Process.* 89, 925–946. doi: 10.1016/S0749-5978(02)00036-5

Sahlins, M. D. (1972). *Stone Age Economics (No. 306.3 S2)*. London: Routledge.

Sharpley, R. (2014). Host perceptions of tourism: a review of the research. *Tour. Manag.* 42, 37–49. doi: 10.1016/j.tourman.2013.10.007

Shore, L. M., and Coyle-Shapiro, J. A.-M. (2003). New developments in the employee-organization relationship. *J. Organ. Behav.* 24, 443–450. doi: 10.1002/job.212

Sinclair, R. R., and Tetrick, L. E. (1995). Social exchange and union commitment: a comparison of union instrumentality and union support perceptions. *J. Organ. Behav.* 16, 669–680. doi: 10.1002/job.4030160706

Tepper, B. J., Carr, J. C., Breaux, D. M., Geider, S., Hu, C., and Hua, W. (2009). Abusive supervision, intentions to quit, and employees' workplace deviance: a power/dependence analysis. *Organ. Behav. Hum. Decis. Process.* 109, 156–167. doi: 10.1016/j.obhdp.2009.03.004

Tepper, B. J., and Taylor, E. C. (2003). Relationships among supervisors' and subordinates' procedural justice perceptions and organizational citizenship behaviors. *Acad. Manag. J.* 46, 97–105. doi: 10.5465/30040679

Thibault, J. W., and Kelley, H. H. (1959). *The Social Psychology of Groups*. New York: John Wiley.

Tsui, A. S., and Wang, D. X. (2002). "Employment relationships from the employer's perspective: current research and future directions" in *International Review of Industrial and Organizational Psychology*. eds. C. L. Cooper and I. T. Robertson (Chichester: Wiley), 77–114.

Turillo, C. J., Folger, R., Lavelle, J. J., Umphress, E. E., and Gee, J. O. (2002). Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organ. Behav. Hum. Decis. Process.* 89, 839–865. doi: 10.1016/S0749-5978(02)00032-8

Uhl-Bien, M., and Maslyn, J. M. (2003). Reciprocity in manager-subordinate relationships: components, configurations, and outcomes. *J. Manag.* 29, 511–532. doi: 10.1016/S0149-2063_03_00023-0

Wang, D., Tsui, A. S., Zhang, Y., and Ma, L. (2003). Employment relationships and firm performance: evidence from an emerging economy. *J. Organ. Behav.* 24, 511–535. doi: 10.1002/job.213

Witt, L. A. (1991). Equal opportunity perceptions and job attitudes. *J. Soc. Psychol.* 131, 431–433. doi: 10.1080/00224545.1991.9713869

Witt, L. A. (1992). Exchange ideology as a moderator of the relationships between the importance of participation in decision making and job attitudes. *Hum. Relat.* 45, 73–85. doi: 10.1177/001872679204500104

Witt, L. A., and Broach, D. (1993). Exchange ideology as a moderator of the procedural justice-satisfaction relationship. *J. Soc. Psychol.* 133, 97–103. doi: 10.1080/00224545.1993.9712122

Witt, L. A., Kacmar, K. M., and Andrews, M. C. (2001). The interactive effects of procedural justice and exchange ideology on supervisor-rated commitment. *J. Organ. Behav.* 22, 505–515. doi: 10.1002/job.99

Witt, L. A., and Wilson, J. W. (1990). Income sufficiency as a predictor of job satisfaction and organizational commitment: dispositional differences. *J. Soc. Psychol.* 130, 267–268. doi: 10.1080/00224545.1990.9924578

Yadav, M. S. (2014). Enhancing theory development in marketing. *AMS Rev.* 4, 1–4. doi: 10.1007/s13162-014-0059-z

Appendix

Appendix 1 Studies examining individual differences in reciprocity.

Year	Author(s)	Exchange orientation
1986	Eisenberger et al. (1986)	Exchange ideology
1987	Eisenberger et al. (1987)	Reciprocation ideology
1990	Witt and Wilson (1990)	Exchange ideology
1991	Witt (1991)	Exchange ideology
1992	Witt (1992)	Exchange ideology
1992	Cotterell et al. (1992)	Reciprocation wariness creditor ideology
1993	Witt and Broach (1993)	Exchange ideology
1994	Orpen (1994)	Exchange ideology
1995	Sinclair and Tetrick (1995)	Exchange ideology
1999	Lynch et al. (1999)	Reciprocation wariness
2000	de Ruyter and Wetzels (2000)	Relational exchange orientation
2000	Ladd and Henry (2000)	Exchange ideology
2001	Witt et al. (2001)	Exchange ideology
2001	Perugini and Gallucci (2001)	Personal norm reciprocity
2001	Eisenberger et al. (2001)	Exchange ideology
2003	Shore and Coyle-Shapiro (2003)	Reciprocity norm acceptance
2003	Gallucci and Perugini (2003)	Personal norm of reciprocity
2003	Andrews et al. (2003)	Exchange ideology
2004	Coyle-Shapiro and Neuman (2004)	Exchange ideology creditor ideology
2004	Eisenberger et al. (2004)	Positive norm of reciprocity Negative norm of reciprocity

All citations in the table are listed in the reference list.

frontiers | Frontiers in Psychology

TYPE Opinion
PUBLISHED 14 February 2023
DOI 10.3389/fpsyg.2023.1025507

Check for updates

OPEN ACCESS

EDITED BY
Guy Dove,
University of Louisville, United States

REVIEWED BY
Selene Arfini,
University of Pavia, Italy

*CORRESPONDENCE
Christopher Kam
✉ ckam060@uottawa.ca

SPECIALTY SECTION
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

RECEIVED 23 August 2022
ACCEPTED 30 January 2023
PUBLISHED 14 February 2023

CITATION
Kam C (2023) Psychoanalytic contributions in distinguishing willful ignorance and rational knowledge avoidance.
Front. Psychol. 14:1025507.
doi: 10.3389/fpsyg.2023.1025507

COPYRIGHT
© 2023 Kam. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Psychoanalytic contributions in distinguishing willful ignorance and rational knowledge avoidance

Christopher Kam*

School of Counselling, Psychotherapy and Spirituality, Faculty of Human Sciences, Saint Paul University, Ottawa, ON, Canada

KEYWORDS

self-deception, willful ignorance, knowledge avoidance, psychoanalysis, ego inflation

Introduction

Discussions on self-deception have occurred in academia for a long time (Deweese-Boyd, 2021). Part of the difficulty in defining self-deception is the puzzling nature of how an individual can seem to be both aware and unaware of tricking oneself at the same time (Lewis, 1996). The Stanford Encyclopedia of Philosophy notes that "self-deception involves a person who seems to acquire and maintain some false belief in the teeth of evidence to the contrary as a consequence of some motivation, and who may display behavior suggesting some awareness of the truth" (Deweese-Boyd, 2021). In the midst of exploring the nature of self-deception, there is an attempt to distinguish between willful ignorance and rational knowledge avoidance since the latter can be useful, arguably beneficial in some cases (Arfini and Magnani, 2021) and can assist in overcoming self-deception. Although the nuanced distinction between willful ignorance and rational knowledge avoidance is interesting and helpful, there are gaps explaining the different psychological processes occurring in one compared to the other. Some of these explanatory gaps are in the unconscious dimensions of the mind. Here, psychoanalytic contributions that outline unconscious processing can offer additional explanatory power in making sense of how rational knowledge avoidance can sometimes be prudent while conceptually contrasting it with the imprudence of willful ignorance.

Willful ignorance vs. rational knowledge avoidance

In the research literature, willful ignorance has an all-encompassing quality of "the general avoidance of situations that let someone aware of certain information, evidence, or knowledge" (Arfini and Magnani, 2021, p. 4). In contrast, knowledge avoidance can be rational and relatively specific in avoiding certain information for particular reasons. Arfini and Magnani (2021) write "we can say that people are willfully ignorant of something when they avoid all circumstances that would allow them to acquire that knowledge, even by accident. Instead, people in a condition of knowledge avoidance do not perform the necessary steps to get a specific piece of information" (p. 3). The latter involves avoiding the knowledge of certain information that may emotionally affect one's judgment or reasoning. People who are engaging in knowledge avoidance are "well aware of which information they are avoiding and why" (p. 4–5). It is similar to the notion of rational ignorance, which happens "When the costs of acquiring knowledge outweigh the benefits of possessing it" (Williams, 2021, p. 7,807).

Psychoanalytic contributions to the distinction

This distinction between willful ignorance and rational knowledge avoidance can benefit from insights from the psychoanalytic tradition of psychology that focuses on unconscious processes. Boag (2017) notes that "a mental process is descriptively unconscious if we are presently unaware of it. For example, a belief would be described as descriptively unconscious if it was believed, without the person currently being aware of having the belief" (p. 2). He writes elsewhere that there is a difference between cognitive science's "cognitive unconscious" and the psychoanalytic "dynamic unconscious" (Boag, 2020). The former has non-motivated obstacles to becoming aware of something (e.g., automated processing that is implicit but not presently available to the mind while no motivated repression is involved). The latter has motivated obstacles preventing awareness (e.g., unconscious repression involving defense mechanisms defending against unprocessed pain) (Solms, 2017, 2018; Boag, 2020).

Empirical research shows significant effects of unconscious influences on conscious processes. For example, emotions can be experienced in brain regions without one's conscious awareness (Brooks et al., 2012), subliminal pictures can prime people to influence their narratives unconsciously (Kawakami and Yoshida, 2014), and unconscious influences can affect behaviors and goals without conscious initiation (Bettiga et al., 2017). In addition, there is support showing the difference between one's explicit self-concept and implicit self-concept and that discrepancies between the two are related to psychological suffering (Bosson et al., 2003; Zeigler-Hill and Terry, 2007; Fabbro et al., 2017). For example, implicit self-concept, measured by reaction times to words or ideas, may or may not relate to a person's more explicitly expressed self-concept (Greenwald et al., 2002). Also, clinical experiences in Short-Term Dynamic Psychotherapy (based on psychodynamic principles), show how the lifting of unconscious defense mechanisms such as repression can result in unconscious content emerging that has been unprocessed (Davanloo, 1987; Abbass et al., 2012; Town et al., 2013; Johansson et al., 2014).

The relevance of ego inflation

Carl Jung, one of the leading psychoanalysts of the twentieth century, defined ego inflation as an unconscious expansion of one's personality beyond its proper limits. When this happens, a person identifies with a persona or archetype that produces an exaggerated sense of one's self-importance and is usually motivated by feelings of inferiority (Jung, 1934–1939, 1963; Schlamm, 2020). This provides psychoanalytic commentary on one version of self-deception where a person has a dispositional tendency to have a positive self-image that is unrealistic (Sackeim, 1983). When an ego is inflated, the person views themself as better than everyone else (Helander and Andersson, 2014). This can lead to the person's ego feeling easily pricked, resulting in retaliation against perceived offenders (Bushman and Baumeister, 1998; Neff, 2011). Ego inflation can be motivated by an inferiority complex resulting from psychic fragments of painful inferiority in the mind that have been split-off from conscious awareness due to previous traumatic influences (Jung, 1970). When this happens, a person can feel inferior with low self-worth in the unconscious parts of the mind, namely the "shadow." Here, inferior traits of character that the individual refuses to acknowledge never fully go away; on the contrary, they are continually trying to thrust themselves onto the person's conscious mind (Jung, 1969). Due to the repression, the person has an indirect but even stronger desire for affirmation that is overcompensating and exaggerated in its hunger. This will lead to a pursuit of ego-inflation, where one may seek flattery from others to feel energized by a socially reinforced shiny persona or a thickly bright self-archetype. This self-archetype contains an image of oneself that seems, in everyday language, perfect in symmetry, flow, crispness, and thickness of being. Here, the ego inflation temporarily drowns out the pain from one's unconscious inferiority complex. Such a process, like an addiction that can increase one's propensity toward episodes of willful ignorance, may lead one to feel ecstatic for a moment but dark, heavy and empty after. This is consistent with research that shows that the inflated ego is unstable (Kernis et al., 1989) and constantly threatened by doubts (Jordan et al., 2005).

An illustration of the integration

All this conceptualization can be tied together in an illustration consistent with research showing that the rise of social media inadvertently encourages ego inflation (Jordan et al., 2014). Pretend there is a young adult named Suzy. Suzy, through an extended journey of self-discovery and awareness of her past patterns, is aware of her propensity to fall victim to a certain type of ego inflation, namely flattery from social media feedback for her online posts. She realizes that this tendency is due to an inferiority complex formed from her past, where she felt low self-worth and tried to compensate for it by manipulating situations to elicit flattering compliments. She realized that she was never satisfied with flattery for long and always needed more. She is aware that in the recent past, she used to be preoccupied with and addicted to her social media account, specifically by checking how many likes, comments, and flattering pieces of feedback she received from her posts. This would lead to a roller coaster of first identifying herself with a shiny, thickly bright persona or self-archetype that others would reinforce through flattery. This led to an exaggerated sense of self-importance, which eventually led to a need for more flattery, eventual jealousy, and then emptiness from comparing herself with others who ended up seeming more "shiny" in "thick brightness" in social media attention. Depressive and empty feelings would follow the dopamine high. Recognizing this propensity, Suzy now tries to be prudent in her use of social media and has decided for the next few months to try an experiment by not checking any notifications of feedback from any of her social media posts. This is an example of rational knowledge avoidance, since it describes "a condition in which agents avoid some knowledge to refrain from anticipated costs (in terms of pain, anxiety, or regret) of possessing it…In these situations, people avoid acquiring those pieces of information that would impact their emotional state, reasoning abilities, and decisions" (Arfini and Magnani, 2021, p. 6). In this case, Suzy is cognizant of her propensity to relapse into a social media addiction that seeks flattery that is fueled by an inner sense of lack or inner deficiency (Naranjo, 1994). Similar illustrations could be made in this framework of using rational knowledge avoidance to prevent ego inflation with other examples of addictions such as substance abuse, gambling, or alcoholism.

To complete the illustration of this article's integrated concepts, it can be noted that if Suzy decided to indulge in flattering social media feedback without any boundaries or self-control, she would engage in ego inflation which would lead down another path, namely willful ignorance. For instance, if this alternative scenario occurred to the extent where she intentionally ignored and disregarded all constructive criticism given in response to her social media posts, then she would be engaging in willful ignorance, specifically in the form of wishful thinking, which is a positive illusion motivated by believing in one's wishes while avoiding content that is inconsistent with it (Sigall et al., 2000; Mayraz, 2011; Jefferson et al., 2017). Here, Suzy would be psychoanalytically motivated to inflate her ego and identify herself as a shiny and thickly bright persona or self-archetype and be motivated to disregard and become ignorant of any disconfirming information that would challenge her ego-inflated view of herself. The motivation of this type of information avoidance would come from an unconscious desire to compensate for one's feelings of inferiority from one's inferiority complex, which would be the opposite of rational knowledge avoidance. This would last as long as the duration of the spell of ego inflation.

From this integrative illustration, we can see that, from a psychoanalytic perspective, rational knowledge avoidance can occur when a person intends to prevent ego inflation while willful ignorance can occur when a person wants to indulge in it.

Conclusion

This article explored the distinction between rational knowledge avoidance and willful ignorance through the lens of psychoanalytic concepts, particularly ego inflation. Relevant conceptual and empirical work were outlined in an integrative way that climaxed with an illustration. Future work can be done to study these integrated concepts from an empirical level and see if they hold water in the laboratory and everyday life. Clinical applications can then be made in terms of when knowledge avoidance can be rational and prudent as well as when it can be imprudent, dysfunctional, and detrimental to wellbeing. As it is harder to regulate ego inflation when individuals are unaware of their own underlying dynamics of themselves, more awareness of these dynamics can help (Yilmaz, 2020). Understanding the difference between willful ignorance and rational knowledge avoidance from integrating philosophical concepts, psychoanalytic commentary, and empirical studies can help individuals understand when it is in their best interest to know something and when it is not.

Author contributions

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

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbass, A., Town, J., and Driessen, E. (2012). Intensive short-term dynamic psychotherapy: a systematic review and metaanalysis of outcome research. *Harv. Rev. Psychiatry* 20, 97–108. doi: 10.3109/10673229.2012.677347

Arfini, S., and Magnani, L. (2021). Embodied irrationality? Knowledge avoidance, willful ignorance, and the paradox of autonomy. *Front. Psychol.* 12, 769591. doi: 10.3389/fpsyg.2021.769591

Bettiga, D., Lamberti, L., and Noci, G. (2017). Do mind and body agree? Unconscious versus conscious arousal in product attitude formation. *J. Bus. Res.* 74, 108–117. doi: 10.1016/j.jbusres.2017.02.008

Boag, S. (2017). "Conscious, preconscious, and unconscious," in *Encyclopedia of Persaonlity and Individual Differences*, eds V. Zeigler-Hill, and T. Shackelford (Cham: Springer), 1–8. doi: 10.1007/978-3-319-28099-8_1370-1

Boag, S. (2020). Reflective awareness, repression, and the cognitive unconscious. *Psychoanalytic Psychology* 37, 18–27. doi: 10.1037/pap0000276

Bosson, J. K., Brown, R. P., Zeigler-Hill, V., and Swann, W. B. Jr. (2003). Self enhancement tendencies among people with high explicit self-esteem: the moderating role of implicit self-esteem. *Self Ident.* 2, 169–187. doi: 10.1080/15298860309029

Brooks, S. J., Savov, V., Allzen, E., Benedict, C., Fredriksson, R., and Schioth, H. B. (2012). Exposure to subliminal arousing stimuli induces robust activation in the amygdala, hippocampus, anterior cingulate, insular cortex and primary visual cortex: a systematic metaanalysis of fMRI studies. *NeuroImage* 59, 2962–2973. doi: 10.1016/j.neuroimage.2011.09.077

Bushman, B. J., and Baumeister, R. F. (1998). Theatened egotism, narcissism, self-esteem, and direct and displaced aggression: does self-love or self-hate lead to violence? *J. Pers. Soc. Psychol.* 75, 219–229. doi: 10.1037/0022-3514.75.1.219

Davanloo, H. (1987). "Unconscious therapeutic alliance," in *Frontiers of Dynamic Psychotherapy*, ed P. Buirski (New York, NY: Brunner Meisel), 64–88.

Deweese-Boyd, I. (2021). *Self-Deception.* Stanford, CA: The Stanford Encyclopedia of Philosophy.

Fabbro, A., Crescentini, C., Matiz, A., Clarici, A., and Fabbro, F. (2017). Effects of mindfulness meditation on conscious and non-conscious components of the mind. *Appl. Sci.* 7, 349–361. doi: 10.3390/app7040349

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farn-Ham, S. D., Nosek, B. A., and Mellot, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychol. Rev.* 109, 3–25. doi: 10.1037/0033-295X.109.1.3

Helander, M., and Andersson, M. (2014). *Inflated Ego or Low Impulse Control: Which Personality Aspect Predicts Juvenile Delinquency Better?* (Unpublished student's thesis). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:oru:diva-33641 (accessed November 11, 2018).

Jefferson, A., Bortolotti, L., and Kuzmanovic, B. (2017). What is unrealistic optimism? *Conscious Cogn.* 50, 3–11. doi: 10.1016/j.concog.2016.10.005

Johansson, R., Town, J. M., and Abbass, A. (2014). Davanloo's intensive short-term dynamic psychotherapy in a tertiary psychotherapy service: overall effectiveness and association between unlocking the unconscious and outcome. *PeerJ* 2, e548. doi: 10.7717/peerj.548

Jordan, C. H., Giacomin, M., and Kopp, L. (2014). Let go of your (inflated) ego: caring more about others reduces narcissistic tendencies. *Soc. Pers. Psychol. Compass* 8, 511–523. doi: 10.1111/spc3.12128

Jordan, C. H., and Spencer, S. J., and Zanna, M. P. (2005). Types of high self-esteem and prejudice: how implicit self-esteem relates to ethnic discrimination among high explicit self-esteem individuals. *Pers. Soc. Psychol. Bull.* 31, 693–702. doi: 10.1177/0146167204271580

Jung, C. G. (1934–1939). *Nietzsche's "Zarathustra": Notes of the Seminar Given in 1934–1939,* ed J. Jarret (London: Routledge).

Jung, C. G. (1963). *Memories, Dreams, Reflections*, ed A. Jaffe (London: Fontana Press).

Jung, C. G. (1969). *Archetypes and the Collective Unconscious, Vol. 9 (Part 1),* ed C. G. Jung (Princeton, NJ: Princeton University Press).

Jung, C. G. (1970). *Structure and Dynamics of the Psyche, Vol. 8*, ed C. G. Jung (Princeton, NJ: Princeton University Press).

Kawakami, N., and Yoshida, F. (2014). Perceiving a story outside of conscious awareness: when we infer narrative attributes from subliminal sequential stimuli. *Conscious. Cogn.* 33, 53–66. doi: 10.1016/j.concog.2014.12.001

Kernis, M. H., Grannemann, B. D., and Barclay, L. C. (1989). Stability and level of self-esteem as predictors of anger arousal and hostility. *J. Pers. Soc. Psychol.* 56, 1013–1022. doi: 10.1037/0022-3514.56.6.1013

Lewis, B. (1996). Self-deception: a postmodern reflection. *J. Theor. Philos. Psychol.* 16, 49–66. doi: 10.1037/h0091152

Mayraz, G. (2011). Wishful thinking. *SSRN Electro. J.* 1955644. doi: 10.2139/ssrn.1955644

Naranjo, C. (1994). *Character and Neurosis: An Integrative View*. Nevada City, CA: Gateways/IDHHB.

Neff, K. D. (2011). Self-compassion, self-esteem, and well-being. *Soc. Pers. Psychol. Compass* 5 1–12. doi: 10.1111/j.1751-9004.2010.00330.x

Sackeim, H. A. (1983). "Self-deception, depression, and self-esteem: the adaptive value of lying to oneself," in *Empirical Studies of Psychoanalytic Theory, Vol. 1*, ed J. Masling (Hillsdale, NJ: Erlbaum), 101–158.

Schlamm, L. (2020). "Enneagram," in *Encyclopedia of Psychology and Religion*, eds D. A. Leeming, K. Madden, and S. Marlan (New York, NY: Springer), 870–873.

Sigall, H., Kruglanski, A., and Fyock, J. (2000). Wishful thinking and procrastination. *J. Soc. Behav. Pers.* 15, 283–296.

Solms, M. (2017). What is "the unconscious," and where is it located in the brain? A neuropsychoanalytic perspective. *Ann. N. Y. Acad. Sci.* 1406, 90–97. doi: 10.1111/nyas.13437

Solms, M. L. (2018). The neurobiological underpinnings of psychoanalytic theory and therapy. *Front. Behav. Neurosci.* 12, 294. doi: 10.3389/fnbeh.2018.00294

Town, J. M., Abbass, A., and Bernier, D. (2013). Effectiveness and cost effectiveness of Davanloo's intensive short-term dynamic psychotherapy: does unlocking the unconscious make a difference? *Am. J. Psychother.* 67, 89–108. doi: 10.1176/appi.psychotherapy.2013.67.1.89

Williams, D. (2021). Motivated ignorance, rationality, and democratic politics. *Synthese* 198, 7807–7827. doi: 10.1007/s11229-020-02549-8

Yilmaz, H. (2020). Possible result of extreme parenting: power of helicopter parenting attitude to predict ego inflation. *Pegem J. Educ. Instruct.* 10, 523–554. doi: 10.14527/pegegog.2020.018

Zeigler-Hill, V., and Terry, C. (2007). Perfectionism and explicit self-esteem: the moderating role of implicit self-esteem. *Self Ident.* 6, 137–153. doi: 10.1080/15298860601118850

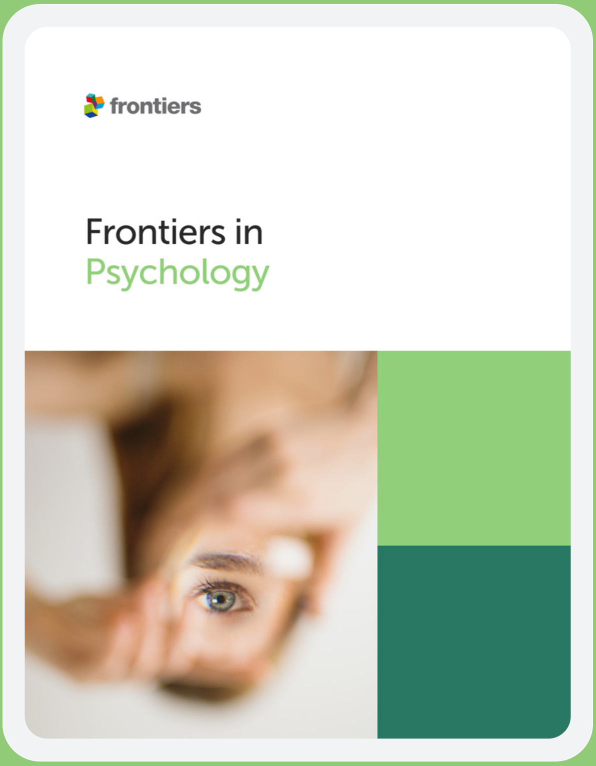
Frontiers in
Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact