BELGRADE PHILOSOPHICAL ANNUAL/
FILOZOFSKI GODIŠNJAK 31/2018
Institute of Philosophy, Faculty of Philosophy,
University of Belgrade
Belgrade, Čika Ljubina 18–20

Belgrade

Year XXXI

YU ISSN 0353-3891

UDK-1

Belgrade Philosophical Annual online at http: //www.f.bg.ac.rs/bpa/index.html

Printed by
Službeni glasnik, Belgrade

This issue is financially supported by
Ministry of Education, Science and Technological Development of the Republic of Serbia

The statement on publication ethics can be found at the journal website
http://www.f.bg.ac.rs/bpa

BELGRADE PHILOSOPHICAL ANNUAL 31/2018

TRENDS IN PHILOSOPHY OF COGNITIVE SCIENCE

Guest editor: Miljana Milojević

TRENDS IN PHILOSOPHY OF
COGNITIVE SCIENCE

Bryce Huebner
Department of Philosophy
Georgetown University

Original Scientific Paper
UDK 165.194:159.955

PICTURING, SIGNIFYING, AND ATTENDING[1]

Abstract. *In this paper, I develop an empirically-driven approach to the relationship between conceptual and non-conceptual representations. I begin by clarifying Wilfrid Sellars's distinction between a non-conceptual capacity to picture significant aspects of our world, and a capacity to stabilize semantic content in the form of conceptual representations that signify those aspects of the world that are relevant to our shared practices. I argue that this distinction helps to clarify the reason why cognition must be understood as embodied and situated. Drawing on recent models of attention and valuation, I then argue that the human brain constructs a dynamic model of the world that it has encountered, encoding higher-level regularities in the form of linguistically structured representations. And I conclude by arguing that this approach to cognition provides a set of critical resources for understanding the situated nature of social cognition.*

Key words: *Situated Cognition; Sellars; Cybernetics; Attention; Social Cognition*

Introduction

Nutmeg is becoming impatient. It is early in the afternoon, and she is starting to feel hungry. She meows and paws at me. So I try to remind her that the auto-feeder will deliver her food at precisely 15:00; but she doesn't seem to listen or care. As you might have guessed, Nutmeg is a cat; so she isn't a language user, and it's unlikely that her thoughts rely on symbolic states with precise conceptual and propositional structure. She seems to have a

[1] The ideas in this paper have been percolating for a very long time. They were first presented at a conference on *Sellars's Legacy,* at the American University of Beirut (26 May 2015). Various parts of this project were then developed in papers that I presented at The University of Maryland, Associazione Italiana di Scienze Cognitive, The George Washington University, Virginia Tech, The NEH Summer Institute on Presupposition and Perception, Minds Online, Universität Stuttgart, Université du Québec à Montréal, The University of Michigan, Georgia State University, and Mind & Life Europe. I am thankful for the discussions I had in each of these places, and for the chance to learn from many amazing philosophers and cognitive scientists. I would also like to offer a special thanks to Rebecca Todd for helping me to understand the importance of implicitly guided attention; I would like to thank Carl Sachs for numerous conversations about Sellars—and many other issues; and I would like to thank Carl Sachs, Ben Elzinga, Dan Williams, and two anonymous referees for this journal for their helpful comments on earlier drafts of this paper. Finally, and most importantly, I would like to thank Ruth Kramer for helping me work through the toughest parts of this paper, and for always being there when I struggle to find a way forward.

doi: 10.5937/BPA1831007H

fairly good sense of where her food will be, and roughly when it will arrive; she also seems to have numerous learned expectations about the micro-world in which she lives; and there are many things that really do seem to matter to her—including access to food and human attention. But so far as I can tell, she doesn't have anything like a mental language that consists of "word-sized concepts, sentence-sized intentional states and argument-sized inferences" (Williams 2018, 153). The neural representations that have been observed in the brains of nonhuman animals typically take the form of topographic maps, representations of local motion, efference copies, and forward models (Thomson & Piccinini 2018). And while such representations are sufficient to guide goal-directed behavior in specific contexts, they offer little insight into the human ability to "impose stability, order, and uniformity upon a conception of the world" (Akins 1996, 368).

While there are notorious difficulties with appeals to conceptual representations (Akins 1996, 367ff; Chomsky 1995; Thompson 2010), psychological explanations of human behavior commonly appeal to symbolic states that are "tailor-made for semantic interpretation" (Egan 2019, 247). This shouldn't be surprising, as such states can be organized to yield networks of computational processes that can "directly encode and exploit the kinds of information that a human agent might consciously access when trying to solve a problem" (Clark 2014, 35; Newell & Simon 1956, 1976). And they seem to provide a nice bridge between our folk-ontology and a computational model of the mind (Schneider 2009; Salisbury & Schneider 2019). Of course, most cognitive scientists and many philosophers acknowledge that appeals to conceptual states are an idealization. But the assumption that human thought relies upon word-shaped concepts and sentence-shaped thoughts persists in many domains, shaping psychological hypotheses, and generating intractable disputes over the nature of mental representation. So we seem to face a dilemma: we can either treat the brain as a computational system, which operates over identifiable neural representations that have little in common with the folk understanding of thought; or we can work to preserve the model of thought that is central to folk-psychology, while abandoning hope when it comes to the computational theory of mind.

There are long running debates about these issues. And I don't intend to address them head on. But I hope to make headway on these issues by providing an empirically-driven approach to the relationship between conceptual and non-conceptual representations; some aspects of this framework will be familiar, but by highlighting the importance of attention and valuation, I show that there is a way of seeing the brain as a dynamic system, which is shaped by our ongoing interactions with the world, and which also encodes higher-level regularities that take the form of linguistically structured representations. In making this case, I begin with the model of cognition that was developed by Wilfrid Sellars (1960). Sellars distinguished two kinds of cognitive capacities: a non-conceptual capacity to *picture* significant aspects of our world; and a capacity to stabilize semantic content in the form of conceptual representations that *signify* those aspects of the world that are relevant to our shared practices. While this may seem like an odd place to start, I contend that the approach that Sellars advances helps to clarify the reason why cognition must be understood as embodied and situated, even if we retain a relatively traditional approach to cognitive processing. From here, I move to the implicit forms of valuation and attention that organize patterns of thought and behavior non-conceptually; and I argue that this addition to Sellars model can help to provide a more contemporary foundation for understanding the nature of conceptual thought. And I conclude by showing how this approach to cognition can be applied to help clarify the situated nature of social cognition.

1. Picturing the world

In one of the earliest philosophical articulations of the computational theory of mind, Sellars (1960) describes a robot that encodes information about the world in memory, by 'printing sentences on its tape'. This robot wanders "around the world, scanning its environment, recording its 'observations', enriching its tape with deductive and inductive 'inferences' from its 'observations' and guiding its 'conduct' by 'practical syllogisms' which apply its wired in 'resolutions' to the circumstances in which it 'finds itself'" (Sellars 1960, §39). Over time, the robot develops a better 'understanding' of the world, and becomes more attuned to the aspects of the world that are relevant to its needs. This robot looks a lot like the kind of learning machine that Alan Turing (1950) posits in the final section of "Computing machinery and intelligence". And at least initially, it seems to rely on symbolic forms of representation, which can be nicely organized into a language of thought. But as the generous use of scare quotes suggests, Sellars was skeptical of this characterization of the robot as a conceptual system; and his distinction between picturing and signifying is intended to provide a way of decoupling the computational features of thought from the capacity for conceptual representation (Sellars 1960, §32).[2] The significance of this claim will become clearer over the course of this paper; but for now, the important thing to note is that Sellars's distinction turns on an account of the embodied strategies that cognitive systems develop as they learn to engage with aspects of the world that matter to them. Sellars argues that cognitive systems develop models of the world that they inhabit using simple forms of error-driven learning; the resulting models are highly structured, but they are not conceptually organized—they are holistically structured models that are constantly

[2] For alternative discussions of picturing, which explore Sellars's claims in more detail, see Levine (2007), O'Shea (2007), and Sachs (2018).

updated in light of new information. That said, they provide the foundation for conceptual thought. But to see how they do so, it is necessary to first clarify the dynamic structure of these models.

1.1 What is a picture?

Throughout his career, Sellars (1956; 1960; 1974; 1981) explored a variety of different ways of understanding capacities for learning and self-regulation. He was an avid reader in the cognitive and behavioral sciences, and he often drew on something like Edward Tolman's *purposive behaviorism* (cf., Olen 2018). Tolman (1932) had argued that learning is an active and goal-directed process: animals explore different strategies for pursuing things that matter to them (e.g., finding food and cuddling with friends); they use their successes and failures to improve their understanding of the world; and over time, they construct cognitive maps that will effectively guide their behavior, at least in familiar contexts (Tolman 1948). Likewise, Sellars (1981 §56) argues that "to be a representational state, a state of an organism must be the manifestation of a system of dispositions and propensities by virtue of which the organism constructs maps of itself in its environment, and locates itself and its behavior on the map"; he claims that cognitive maps play a critical role in the guidance of flexible and adaptive forms of behavior; and, he appeals to forms of feedback-driven learning to explain how we construct "an increasingly adequate and detailed picture of" the world (Sellars 1960, §40). But Tolman never explained how cognitive maps were realized in the brain; and at points, he seems to treat them as internal images that show up to the animal who uses them. By contrast, Sellars argues that cybernetic theory throws "light on the way that cerebral patterns and dispositions picture the world" (Sellars 1960, §59). And this is where his approach diverges from standard forms of machine functionalism, yielding an embodied and situated understanding of thought and agency. Put much too simply for now, a cybernetic approach to cognition draws our attention to the control of purposive behavior; it highlights the importance of ongoing feedback in the production of resilient dynamic relations between an organism and the world; and while cybernetic systems can be designed to accommodate symbolic representations and person-level inferences, they tend to be more concerned with control over the actions that allow a system to survive and flourish in its natural environment.

While Sellars doesn't go into detail on any of these points, they do play a prominent role in his discussion of embodiment (as I argue in the next sub-section). And they do seem to be implicit in the accounts of cognition that he draws upon. Specifically, his reference to cybernetic theory as it would have been understood in the 1960 suggests a view of picturing that depends on informational relations, which can be implemented in the network structure of a brain. In the early 1940s, Warren McCulloch & Walter Pitts (1943) argued that the all-or-nothing character of neural activity allowed individual neurons to function as logic gates, which could be cyclically organized to represent logically structured propositions. Skeptics quickly noted that neural activity could not be captured by digital flows of propositional states (Abraham 2019); and a variety of cybernetic alternatives rapidly emerged, focusing on the nature of behavioral control in animals and machines (Wiener 1948). These approaches retained the commitment to mechanisms that detected, processed, interpreted, and stored information at multiple points in the brain. For example, Donald Hebb (1949) argued that associative learning could be implemented in a system that relied on the strengthening of connections between neurons that were active in concert, inducing lasting cellular changes that could facilitate the storage of information. Frank Rosenblatt (1958) used feedback-driven learning to show how classificatory information could be acquired and stored in the dynamic connections within a neural network. And having learned that cells are responsive to specific features of objects (e.g., length, orientation, contrast), Oliver Selfridge (1958) developed a pandemonium architecture, which used multiple 'demons', working in parallel to extract meaningful patterns from noisy signals.

As Sellars (1981, §64) notes, however, the storage and processing of information is only part of the story when it comes to the nature of mental representation: the cognitive maps we construct and use exploit a network of interconnected states, which drive motor activity, and generate reliable strategies for getting around in the world. And by the late 1950s, a similar insight had inspired research into the frog's capacity to reliably detect, track, and consume fast moving insects. Building on the insights that had led Selfridge to develop the Pandemonium architecture, Jerome Lettvin and his colleagues (1959) argued that a plausible understanding of mental representations should appeal to the role of interacting systems in the guidance of goal-directed behavior. But they moved beyond computational models, to show how such operations could be implemented in a biological brain. They identified cells that were responsive to small dark shapes moving across the visual field; they suggested that computations carried out by these cells could be specified in terms of operations over edges and contrasts, curvature, movement, and luminance; and they argued that information from these cells was stored in a retinotopically organized map in the superior colliculus, which facilitated control over food-seeking behavior.

The representations posited by this model were deeply tied to adaptive behavior, and they were characterized in terms of feature detectors, retinotopic maps, and mathematical functions. Moreover, the systems that Lettvin and his colleagues posited were tightly coupled to the flow of information through specific neural systems, yielding a highly promising approach to linking neuroscience and behavior (Maturana et al 1959; Lettvin et al 1959). But just as importantly, these approaches described the relevant class of computations in mechanical terms; and any heuristic gloss of what the frog was representing (e.g., flies, or fast moving insects) would necessarily go beyond what the data could support (cf., Egan 2014). Consequently, while these models revealed important facts about how frogs represent the world, they didn't reveal anything about their ability to track the category 'insect' (this is the primary sense in which they are non-conceptual systems). Indeed, the mechanisms that were discovered in their eyes and brains only seemed to track differences in motion and light. Moreover, these models showed that the ability to reliably track these kinds of differences could be explained without attributing any sensitivity to any conceptual categories to the frog itself.

1.2 Recordings and embodiment

Sellars never addresses the frog's capacity to picture the world. But even if he was unaware of this research, he was concerned with the kind of informational isomorphism that we find in the states of a neural network, and the mechanisms in the superior colliculus that preserve features of the world that are most salient to behavioral guidance. To see why, consider an analogy that he offers to the way that the groove on a vinyl record pictures a musical performance (Sellars 1960, §40). The groove is caused by the acoustic properties of a specific performance; and it records the sonic profile of the performance, using a function that maps acoustic differences onto differences in groove depth; and this recording can be recovered, using an inverting function to map differences in groove depth onto a sonic output. But while the record is always machine-readable, it only becomes person-usable when it is placed on a turntable, with the right kind of needle, rotating at the right speed, and sending the right kind of signal through an amplifier to a set of speakers. As I read him, this is why Sellars (1960, §40) suggests that this picture "cannot be abstracted from the procedures involved in making and playing the record".[3]

Building on this analogy, we can see cognitive maps as recordings, which are produced by a mechanism that maps perceptible differences in evaluative salience onto differences in the structure of a neural network. In picturing the world, a cognitive system takes in perceptual information, represents it using an isomorphic relation, and uses this representation to guide its purposive

[3] Compare Frances Egan's (2014, 116) description of the computation of the addition function: "A physical system computes the addition function just in case there exists a mapping from physical state types to numbers, such that physical state types related by a causal state-transition relation (p1, p2, p3) are mapped to numbers n, m, and n+m related as addends and sums. Whenever the system goes into the physical state specified under the mapping as n, and then goes into the physical state specified under the mapping as m, it is caused to go into the physical state specified under the mapping as n+m." Importantly, this way of characterizing computation requires nothing more, and nothing less than: 1) a functional mechanism that can process variables that change states (e.g. patterns of neural spiking), 2) in accordance with rules that map inputs to outputs, 3) in ways that are sensitive to the properties of these variables, and to differences between different portions of them (Piccinini 2015; Piccinini & Bahar 2013; Piccinini & Scarantino 2011).

behavior. But like the grooves on a vinyl record, the patterns in a neural network do not picture something just by being present in the structure of a system; picturing "involves the manner in which the patterns...are added to, scanned, and responded to by the other components" of the system (Sellars 1960, §40). And the isomorphic relations that encode information about the world only become a way of picturing things "by virtue of the physical habitus of the" system, that is "by virtue of mechanical and electronic propensities which are rooted, ultimately, in its wiring diagram" (Sellars 1960, §40).

This may seem like a strange way to phrase this claim. After all, the term 'habitus' is typically associated with the sociological research of Pierre Bourdieu. And it is commonly used to identify the network of embodied dispositions that organize an individual's perception of, and actions within, the social world. But Sellars's use of the term derives from the work of Thomas Aquinas.[4] And here, an agent's *habitus* is understood as a mode of being, which arises through their intentional and purposeful use of an extrinsic thing, in a way that "actualizes one of the open-ended range of potentialities for engaging with the world engendered by human reason" (Spencer 2015, 121). For example, my use of the *rakweh* that I bought years ago in Beirut actualizes my ability to make coffee in a specific way; and it does so because my (learned) capacities for coffee-making fit with the function of the *rakweh*, in a way that actualizes a specific range of coffee-making activities. Extending these claims more broadly, we might say that picturing the world actualizes the ability to act in specific ways; and that it does so because we possess learned capacities that fit with the world, and that actualize specific forms of rationally structured action. To acquire a way of picturing the world is thus to become attuned to particular aspects of the world, and to orient our action toward them; but just as importantly, this structure of attunement is an embodied and situated strategy for engaging with the world, which constrains the sources of information that an agent will track and respond to, while delimiting their possibilities for acting.[5]

The mechanisms that produce such pictures must facilitate attunement to salient aspects of the world. They must produce internal states, which are isomorphic to salient features of the world; and in virtue of this fact, they must be able to guide purposive behavior. Finally, keeping with Sellars's appeal to cybernetics, we should see these mechanisms as operating by changing connection weights between neurons, altering their tonic or phasic firing rates, or stabilizing isomorphic relations between the states of a brain and state of the world. And this brings us to Sellars's core insight: picturing

[4] "Being and being known" was originally published in the Proceedings of the American Catholic Philosophical Association, and Sellars claims that it is an exploration of "the profound truth contained in the Thomistic thesis that the senses in their way and the intellect in its way are informed by the natures of external objects and events".

[5] For a far more compelling defense of a nearby perspective, see Kukla (2017).

is an *activity*, which depends on the use of a cognitive map to actualize one of an agent's capacities for acting in an embodied, embedded, and situated way; and since cognitive maps provide a representation of an agent's place in the world that *they have encountered,* they will always be deeply tied to that agent's capacities for action (cf., Sellars 1978 §28–29).[6] This insight also motivated much of the initial research in cybernetic theory. As Kenneth Craik (1943, 61) famously puts this point, if a cognitive system "carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way react in a much fuller, safer, and more competent manner to the emergencies which face it". But at the same time, such models must be open-ended and revisable. They cannot rely on static symbolic representations, and they cannot be decoupled from the world in which they are embedded. And by focusing on these aspects of Sellars's model, we come to a more contemporary account of picturing.

2. A neo-Sellarsian approach to picturing

Thus far, I have argued that picturing is an activity that depends on the use of a cognitive map to actualize specific capacities for acting. In acquiring a way of picturing the world, an agent becomes attuned to salient phenomena, in ways that yield embodied and situated strategies for action, constrain the sources of information that an agent tracks and respond to, and delimit an agent's possibilities for acting. But at the same time, such capacities must be flexible, open-ended, and revisable. This is something that Sellars seems to have recognized, but in thinking about the nature of picturing as a foundation for nonconceptual thought, the salience of this flexibility is crucially important. My aim in this section is to show that recent research on evaluative learning and the implicit guidance of attention provide a way of thinking about how these capacities emerge and stabilize in human minds. And while the underlying processes aren't exactly maps or pictures, they do record information about the world in patterns of neural connectivity, and they sustain the kinds of behavioral capacities that I have been discussing. So they capture the core aspects of Sellars's theory of picturing. That said, my discussion of attention and valuation go beyond anything that Sellars ever said. And they draw us closer to redeeming these cognitive capacities in the coin on cybernetic and neuroscientific data.

[6] Here, too, Sellars seems to be following Tolman (1948, 192), who argues that mind is "more like a map control room than it is like an old fashioned telephone exchange. The stimuli, which are allowed in, are not connected by just simple one-to-one switches to the outgoing response. Rather, the incoming impulses are usually worked over and elaborated in the central control room into a tentative, cognitive-like map of the environment. And it is this tentative map, indicating routes and paths and environmental relationships, which finally determines what responses, if any, the animal will finally release."

There is a broad consensus that typical humans rely upon a network of interacting systems—including the basal ganglia, amygdala, and anterior insula—to evaluate multiple kinds of information in parallel (cf., Gardner et al submitted; Gershman in press). Some of these systems track fluctuations in the value of rewards, others monitor changes in the probability of gains and losses, and others adjust subjective estimates of risk and uncertainty in light of experienced feedback (Montague et al 2012; Adolphs 2010). For example, the tonic and phasing spiking activity of dopaminergic neurons in the basal ganglia represent the predicted value and distribution of rewards; these patterns of activity are adjusted when rewards are better or worse than expected; and these adjustments continue until the value and distribution of rewards are predicted accurately (Schultz, 1998, 2010). Similar systems in the amygdala and striatum generate and revise predictions about aversive stimuli (Delgado et al 2008). And recent data suggest that subsecond dopamine fluctuations in the human striatum facilitate the computation of expected as well as merely possible outcomes (Kishida et al 2016). These systems sustain patterns of attunement to the evaluatively salient aspects of the world through Pavlovian and classical conditioning; their behavior is best understood computationally, but they do not rely on symbolic representations with determinate conceptual content. That said, they do play a critical role in orienting behavior toward evaluatively significant phenomena. And so long as the rate and value of risks, rewards, and threats remain relatively stable, these systems will tend to produce viable capacities to evaluate current and counterfactual situations (Railton 2014).

Like the cutting of grooves into a vinyl record, the activity of these computational systems shapes patterns of neural response, in ways that can be mapped onto properties of the world (e.g., reward, threat, and risk). While it would be a mistake to think that these systems yield anything like a geometric or relational map of the world, they do facilitate the storage of evaluatively salient information. They allow typical humans and many other animals to represent the rewards, threats, and risks they have previously encountered. And they orient patterns of thought and behavior, by constraining the sources of information that a typical human will track and respond to (for reviews in different domains, see Crockett 2013; Cushman 2013; Haas 2017; Huebner 2016). Importantly, these systems also play a critical role in the guidance of implicit attention (Anderson 2016; Todd and Manaligod 2017). And recent approaches to attention have claimed that they are integrated into a priority map of an attentional landscape, which is shaped by ongoing competition for selection between networks of value-driven, goal-directed, and stimulus-driven influences (Anderson 2016, 32; Awh et al 2012).

Many forms of attention are fast and automatic, and they appear to be shaped by lingering biases of selection history, especially by past experiences of reward (Theeuwes 2018). And according to one plausible hypothesis about how this occurs, dopaminergic signals enhance the perception of evaluatively salient information in sensory cortex, as sensory states that are predictive of reward are prioritized across situations, sensory modalities, and stimulus properties—including locations, complex objects and their properties, and characteristics of experienced scenes. On this model, the resulting enhancement of cortical signaling then biases competition between sensory inputs, in ways that affect the salience of things in the world. This hypothesis gains support from the fact that signals from the reward system do seem to influence attention in ways that reflect "formal models of reward learning, including reflecting a common neural currency for value that scales with the amount of reward available in the current task" (Anderson 2016, 35). But these data are also consistent with the hypothesis that the reward system estimates the current state of the animal and its environment, and shapes the interpretation of sensory information in ways that produce patterns of perceptual orienting and behavioral response (Krauzlis et al 2014). According to this approach, attention is an effect of the reward system attempting to interpret sensory data, in light of prior knowledge, and the agent's current state. And intriguingly, connections between the superior colliculus and the basal ganglia provide a plausible point where information from sensory systems might be integrated with evaluative information. Finally, there is some reason to believe that a priority map is realized by a network of retinotopically organized structures in the parietal cortex, allowing an agent to "represent the task relevance, learned value, and physical salience of stimuli" in a single representational structure (Anderson 2016, 32).

Each of these perspectives suggests that strategies for processing current information will often reflect past experiences of reward, yielding habitual forms of attention that are anchored to learned patterns of evaluation (Anderson 2016, 35). However, we are always in some kind of affective and evaluative state, no matter what is happening, and no matter how neutral we feel (Lindquist 2013, 361). And this matters since the evaluative state we are in when we encounter a risk, a threat, or a reward will have a significant impact on how we interpret its salience, and in what information we encode for future action. Hungry people tend to be more attentive to the food that they see, and this enhances their food-related memories (Talmi et al 2013); people who are nervous about snakes tend to see them everywhere while hiking on a trail (Machery 2016); and the experience of economic precarity tends to enhance the salience of economic information (Shah et al 2018), and perhaps even shape the way that people perceive race (Krosch & Amodio 2014). These effects are important, but not just because of the shifts in what people perceive in the moment. To the extent that these tendencies shape what people notice, what they remember, and what they imagine to be possible, they will impact the structure of the picture that someone forms of the world (cf., De Brigard et al 2017, 2018). Put somewhat differently, such effects are part of the history that implicitly guides attention, and that determines how an agent will behave in any future context. And this will be true even if it turns out that perceptual systems have relatively stable, and relatively modular properties (Gross 2017; Machery 2016).

To explain these more pervasive kinds of effects, Rebecca Todd and her colleagues have developed a model of attention that posits priority maps, which are shaped by numerous aspects of an agent's 'history'. (Kryklyvyy & Todd 2018; Todd & Manaligod 2017) According to this model, statistically stable features of an agent's environment will tend to be prioritized, independently of their intrinsic salience or goal relevance (Zhao et al 2013); and this is because people are learning to map the structure of the world that they encounter, by adjusting their response to attentionally salient phenomena. Likewise, reward-driven learning helps to guide attention toward rewarding stimuli, and away from aversive stimuli; this helps an agent orient their behavior within a highly structured evaluative landscape. And in some cases, attentional landscapes can become attuned to the co-occurrence of related features or objects, in ways that simplify the pursuit of complex tasks and goals (Todd & Manaligod 2017, 123). The crucial idea, here, is two-fold: the structure of an agent's priority map is shaped by their history, broadly construed; and the topography of this map determines which features of the world are likely to attract or repel attention. This is why combat veterans returning from Afghanistan tend to prioritize combat-related stimuli (Todd et al., 2015b); it is why passengers on a flight that barely avoided crashing in the Atlantic Ocean continue to show "attentional tuning to stimuli associated with the crash years after the event" (Lee et al 2013; cited in Todd & Manaligod 2017, 124); and more mundanely, it is why interpersonal relationships often fall into habitual patterns of misunderstanding, which are grounded in past interactions and their evaluative salience. In each case, attentional parameters are adjusted to yield a stable picture of the world, which highlights the aspects of an environment that are likely to be relevant to current and ongoing projects; this picture of the world then shapes both task-based and feature-based forms of attentional salience (Kryklyvyy & Todd 2018).

All told, the implementation of priority maps is likely to be quite complex, as attentional salience is shaped by numerous interacting processes, operating over different properties of the world, and different time-scales. Some forms of attentional mapping may be implemented by geometric or relational maps in the medial temporal lobe, the parietal cortex, or the superior colliculus. But in some cases, visual activity might be modulated more directly by circuits in the amygdala and the locus coeruleus, on the basis of implicit attentional sets that are attuned to affectively and motivationally salient stimuli (Todd et al., 2012; Todd & Manaligod 2017, 127). Given its sensitivity to contextual information, norepinephrine circuits in the locus corellicus are also likely to "play a role in contextualizing the sources of salience that are prioritized in any given state space" (Todd & Manaligod 2017, 128); this hypothesis gains support from data showing that a common genetic variation that affects the availability of norepinephrine (the deletion variant of *ADRA2b*) profoundly impacts the experience of affectively salient stimuli, leading to more vivid emotional experience (Todd et al 2015a). Some of these processes operate rapidly and directly on neural structures in the early visual cortex, biasing competition for selection in favor of things that are affectively or motivationally salient; others operate over longer time-scales, shaping the consolidation of information, as well as the subcortical mechanisms that guide our behavior. But collectively they shape a priority map, which orients attention, and structures patterns of thought and behavior.

Finally, Todd and her colleagues argue that genetic predispositions can shape a priority map by modulating the attentional salience of affectively salient stimuli (Todd et al 2013). Just as importantly, they acknowledge that early experience can make certain phenomena seem more salient, as can repeated engagements with specific stimuli—and these facts become highly salient when we look at patterns of attention that emerge outside of the lab, as I argue in Section 4. But even in these cases, an agent's priority map of the world will be continuously shaped by their actions, and by the situations where they find themselves (Todd & Manaligod 2017, 133). For example, individual commitments can restructure the topography of a priority map, allowing a person to orient toward politically salient information, and to avoid information that contradicts their current goals and values (Whitman et al 2018). Explicitly represented goals can shape what we pay attention to for a short time, though they will always compete with the more implicit forms of habitual attention (Jiang 2017). And the topography of a priority map can even shift to align with our current needs and interests—which are shaped by our memories, and realized by distributed patterns of neural reactivation (Todd & Manaligod 2017, 123). This is important, as the things that should be salient to us when we walk around an unfamiliar city are different from the things that should be salient to us when we walk along a familiar path to the office or a café. But at the same time, many of our attentional strategies will drift toward statistically stable and evaluatively significant aspects of the world as we encounter it, even if different attentional sets will be highlighted in different contexts.

The upshot of this neo-Sellarsian account of picturing is that we possess non-conceptual, but highly structured priority maps, which shape our patterns of attention, and which guide goal directed behavior.[7] In line with Sellars's argument, I have suggested that our most basic ways of tracking and responding to the structure of our environment are implemented by priority maps, which are constructed through a process that relies on some endogenous constraints, as well as some forms of reward-driven and statistical learning. These mechanisms operate mechanically, through feedback-driven learning; and they are shaped by a person's practical engagements with the world. Consequently, they yield embodied and situated capacities to act in

[7] I contend that these sub-personal processes, which operate through patterns of attunement, are unlikely to carve the world in ways that map the categorical structure of conceptual thought. For further discussions of this complex issue see Huebner (in press), which includes a further discussion of the example that I explore in the next section.

ways that are highly responsive to regularities in the environment, and they do so without relying on conceptually structured thoughts (thereby paralleling Nutmeg's ability to act responsively, even though she lacks complex beliefs about the world). But humans live in linguistically structured environments; and their priority maps can represent semantically intelligible contents. These representations will often become relatively stable aspects of a person's way of representing the world, which reflect their ability to describe things in conceptual terms. And these conceptually structured thoughts will often shape the way that typically functioning people encounter the world. To understand how these kinds of capacities arise, we need to turn to a more robust theory of signifying, which explains how word-and-sentence shaped thoughts are implemented within the framework of non-conceptual picturing that I have developed thus far (cf., Levine 2007, 254).

3. How do concepts signify?

In contrast to the standard assumption that conceptual states acquire the content that they have through causal contact with the world, Sellars argues that what a term signifies is a matter of functional classification, which is situated within a broader network of social and historical practices. He contends that learning what a word signifies is a matter of learning how to use that word in various contexts, in ways that accord with local inferential practices; though he also acknowledges that complex networks of causal relations sustain "linguistic behavior both in its own internal patterns and in its relationship to entities in the world" (O'Shea 2007). Though Sellars never puts the point exactly this way, we can treat meaningful uses of language as part of a typical person's *habitus,* and we can treat signifying as an activity that depends on the use of parts of a cognitive map to actualize capacities for rational thought and behavior. This is an interesting claim, which pushes us toward a novel way of understanding conceptual content. And to see what this amounts to, it will help to first examine how meaningful representations operate across different languages, before returning to claims about the nature mental representation.[8]

[8] How is the approach I develop related to more familiar discussions of semantics? To begin with, I hold that there are formal and logical constraints on sentence formation, which are acquired by way of a socially situated learning process. So I accept a form of realism and internalism about formal semantics, where meaning is constrained by hierarchical structures that are interpretable by a system that implements what linguists typically call Logical Form (LF). But we cannot infer much about the worldview of a speaker from the grammar of their language (Hale 1986). So I contend that the resulting logically-structured expressions will always be underspecified, and that hearers will have to infer an understanding of the similarities and differences between their picture of the world and the speaker's. This requires adopting an interpretivist lexical semantics, where meaning is a matter of functional classification. Like Sellars, I thus hold that assumptions about meaning will tend to track similarities in the ways that linguistic terms are used

When I learn from an Arabic speaking friend that ‘ةوهق (*qahwah*) means coffee’, I learn that their use of ‘ةوهق’ plays the same functional role as my use of ‘coffee’. I don’t learn anything new about the meaning of ‘coffee’. But I do gain a sense of how they are likely to use ‘ةوهق’. Of course, there will be numerous differences in the ways that we write and speak the relevant word; and there will be differences in the ways that we picture the relevant beverage. But in learning that ‘ةوهق means coffee’, we open up space to discuss salient aspects of the delicious beverage, and to exchange reasons for including or excluding more marginal substances (e.g., Is Dunkin’ Donuts cold brew really coffee? How about a shot of Starbucks espresso in 250 ml of milk?). Given our unique social and historical situations, we will each possess a picture of the world that has been partly caused by encounters with coffee. And in general, there will be structural relations between the parts of our pictures which characterize the space of possible coffee experiences. But we can’t access these aspects of our priority maps directly, any more than we could access the recording on a vinyl record without a turntable. Nonetheless, a long conversation would allow us to explore a wide range of inferences that we were both willing to assent to with regard to ‘coffee’; and a shorter discussion might reveal many shared ways of picturing coffee.

The structure of a person’s picture of the world will be at least partially reflected in their linguistic behavior, since there is a tight causal connection between picturing and behaving. And in some cases, the mapping between neural activity and linguistic behavior can even underwrite ways of treating patterns of neural activity “as symbols which have meaning, which belong to the order of signification” (Sellars 1960, §44). But doing so is never straightforward. After all, a person’s picture of coffee-relevant phenomena will shift and develop as they move around the world, learning more about the use of ‘coffee’, and more about the relevant substance. Moreover, differences in the coffee-relevant parts of a cognitive map will emerge as the result of differences in people’s learning histories, as well as difference in the evaluative salience of coffee. So some people will come to represent coffee (still non-conceptually) in a highly abstract way, as the black stuff that they tend to drink in the morning; while others will develop a more detailed picture, which is structured around knowledge of coffee beans, differences in terroir, and differences in techniques for roasting beans and extracting coffee from them.[9] But just as importantly, we should find that a person’s use of ‘coffee’

(with usage being determined by something like language-entry rules, intralinguistic transitions, and language-exit rules). To unify these two perspectives, we need something very much like the account of semantic content advanced by Donald Davidson (1986). I currently believe that a promising explanation of how we interpret meaningful claims is likely to approximate Elin McCready’s formal model of emotive meaning and dog whistles, which demonstrates how speakers coordinate to extract meaning from underspecified representations (McCready 2012; Henderson & McCready in press). Over the course of this section, I attempt to unpack this complex set of claims.

[9] Following Tolman (1948, 193), we might say that some maps of the coffee domain are narrow and strip-like, while others are broad and comprehensive. Both types of maps can

can expand or contract to fit different contexts, just as their priority map can change across different contexts. I am likely to be more discriminating when I am talking to friends who are baristas, or coffee connoisseurs; and I am likely to become less discerning as I spend more time with people who couldn't care less about the caffeine containing beverage that they drink. But in any case, the process of learning, storing, and using the parts of pictures for particular purposes can open up the possibility of treating those mental states that are used in the way that I typically use the word 'coffee'; and this is what underwrites the claim that a particular pattern of neural activity signifies 'coffee' (Sellars 1960, §52).

The critical upshot of this discussion is that the dynamic nature of attentional landscapes, and our strategies for picturing the world, make it unlikely that we will find a stable and systematic mental language, though points of stability are likely to emerge in attentional landscapes as people learn to speak a language.[10] Put much too simply, learning to use parts of non-conceptual pictures for thinking, planning, and remembering, requires learned and habitual tendencies to use 'coffee' to label the relevant parts of a picture (Clark 1996). When we turn to questions about mental representation, we must therefore distinguish: 1) the ability to use part of a picture to orient toward coffee-relevant aspects of the world from 2) the ability to label part of a picture with the relevant linguistic concept.[11] Both processes are implemented mechanically. And both are necessary to sustain the functional mapping between the current state of a system, and the everyday use of a linguistic term. This means that conceptual thought is stabilized through the construction and use of a picture of the world, which interacts with our ability to treat aspects of a picture as a meaningful representation of the world, and to use this labelled representation for thought and communication. But it is worth dwelling on these points, as it's easy to miss their importance.

Recall that picturing is a non-conceptual capacity, which orients us toward particular aspects of the world, and serves as the model that we can employ in planning and deliberating. Within such a model, every coffee-

be correct, as far as they go; and both can be used to guide relevant forms of actions. But broad and comprehensive maps allow for a wider range of inferential relations, about a wider range of different substances. Importantly, people with different coffee-maps will be able to have a discussion of coffee, even where there are robust differences in the ways that they represent coffee.

[10] This is not an anti-nativist commitment, nor even an argument against Universal Grammar. While I will not defend this claim here, this approach leaves room for a faculty of language that uses statistical learning to produce stable and resilient linguistic properties (see Lightfoot 2017 for a readable defense of this claim).

[11] I initially developed this claim in the context of a paper on racial bias, where I approached this link as a difference between the capacities that we possess socially, and the capacities that we possess individually. I no longer think that this is the right way to develop the argument. But for another way of thinking about this distinction, which draws on a predictive coding framework, and which may be closer to Sellars's own commitments, see Sachs (2018).

representation depends on coffee-relevant dispositions, which are sustained by reliable tendencies to picture the world in particular ways (Levine 2007, 254). And the aspects of a picture that are coffee-relevant can change over time, as new experiences become relevant, and as old ones become irrelevant. In my own case, 'coffee' thoughts are organized around a network of specific interests (e.g., a desire for caffeine, a desire for a particular taste, and my enjoyment of specific flavor profiles), which orient me toward the construction of a cognitive map that will lead me to pursue, brew, drink, and think about coffee across a wide range of different contexts. These interests allow me to ignore features of coffee-drinking experiences that are irrelevant to future coffee-seeking behavior. And they allow me to orient toward those features of coffee drinking that improve my understanding of the diversity and complexity of the coffee domain. When I try a bean from a new roastery, or from a new region, I may focus on the flavor profile of the coffee that I drink. And when I visit a new city, I will seek out the most highly recommended places to drink coffee.

Over time, the coffee-relevant aspects of my picture of the world have become relatively stable, as I have habituated to thinking about coffee in particular ways, which accord with my historical and social situation. To the extent that I am more snobbish about coffee than many of my friends, this is the result of encounters that have shaped the way that I picture the role of coffee in the world; and my idiosyncratic way of being attuned to the world provides me with the foundation for my conceptually structured thoughts about coffee. But for Sellars, the storing of a representation can only ever be part of the story. And a plausible account of conceptual states will also need to explain how stored information is "added to, scanned, and responded to by the other components" of the cognitive system (Sellars 1960, §40). So it is just as important to note that most typical humans have the capacity to use linguistic labels to identify aspects of their cognitive map, and to re-identify the things they have learned about. There is some reason to suppose that a neural network that processes linguistic information will recapitulate the constituent structure of linguistic representations.[12] Where a person internalizes linguistic practices, this will yield parts of the cognitive system that are less contextually variable, and this will allow for re-usable 'words' that retain the same structural relations across sentences where they occur (NB: we are in the domain of formal semantics here, and not the domain of lexical semantics). To the extent that the interfaces between this system and the mapping system allow for (at least) momentary points of stability, we will find internal structures that can serve as the targets of signification.

[12] Schonbein (2012) defends the relevant conceptual claim by reference to connectionist systems; and Ding et al (2017) show how neural oscillations might sustain the kinds of hierarchical structures that are commonly posited by theories of generative grammar.

These internal structures are likely to remain somewhat idiosyncratic, but they will be shareable, and it will be possible to explore and revise them, precisely because of the way that they integrate stable structure with dynamic contents. So the mapping between a word like ‘coffee’ and the internal state that signifies ‘coffee’ will always be complex. But this shouldn’t be surprising. Even in the linguistic domain, the mapping between conceptual states is often more complex than the matching of two similar words. The differences in orthography between an Arabic and English word might seem salient to a first time observer; but these languages use a single morpheme to signify ‘coffee’, and the English term is descended from ‘قهوة’. But a conceptual understanding of ‘coffee’ doesn’t require using a mono-morphemic term, and the term that is used doesn’t need to be a descendant of ‘قهوة’. For example, an Ojibwe speaker might use ‘makade-mashkikiwaaboo’ to signify ‘coffee’. And while a more literal translation might be something like ‘black medicine water’, the multi-morphemic structure of this term is unlikely to have a deep effect on their conceptual understanding of the relevant substance. There will be differences in the linguistic structures that they construct; and my use of the mono-morphemic lexicalization ‘coffee’ my block my ability to use a phrase like ‘black medicine water’ to express my concept (Poser 1992). But these effects are likely to be generated by the way that syntactic information is stored in the brain; and across many different contexts, we will find that the use of ‘makade-mashkikiwaaboo’ is similar to may use of ‘coffee’. And this is all that must be the case for us to establish a relation of signifying between these two differently structured terms. Pushing further, we might even imagine someone who uses a more complex syntactic construction in the way that I use ‘coffee’. While there would be strange lexical implications, someone could even use ‘amazing and delicious black energy water of the gods’ to signify ‘coffee’, so long as they reliably used this construction in ways that were conceptually similar to my use of ‘coffee’. So long as I can find a way to map my use of ‘coffee’ onto their use of ‘amazing and delicious black energy water of the gods’, I can establish similarities between their picture of the world and mine.[13]

This brings us back to the core Sellarsian claim: There are multiple causal routes to the acquisition and use of a term that signifies ‘coffee’; and the meaning of ‘coffee’ can’t be specified by appeal to any of them! ‘Coffee’ means what it does in virtue of its use in practices of functional categorization. For some people, ‘coffee’ may be little more than a linguistic label, which they have linked to a substance they have read about in books or seen in films; for others the inferential structure of ‘coffee’ may be connected to many

[13] I take this to be a plausible way of redeeming Sellars’s claims about dot-quotation in terms of linguistic mappings. Thanks to Carl Sachs for noticing that this is what I was doing, and for suggesting that I leave further development of this argument for a subsequent paper. As the previous footnote suggests, there is a lot of work to do here.

experiences of drinking, brewing, or growing coffee, which structure their sense of what coffee is, and what coffee can do. And while this leads to patterns of difference in the use of 'coffee' (or whatever term signifies 'coffee'), there will often be enough similarity between two people to sustain practices of giving and asking for reasons. From here, ongoing forms of learning and behavior-shaping can draw people closer to one another, both in the way that they picture the world, and in the ways that they use specific terms for thought and communication. There will always be differences in the class of inferences that two people will draw about 'coffee'; but to the extent that these differences are situated among a broader range of similarities, the presence of such states and capacities will allow for practices of communication that can revise and reshape a person's picture of the world.

4. Socialized attention and distorted maps

My primary claim so far is that people develop strategies for navigating their social world by learning to track the things that are most salient to them. This has important implications for the way that they attend to different features of the world; and it has important implications for the ways that they learn to categorize, and to think conceptually. On the one hand, people "learn to navigate the world by attending to the predictability and frequency of objects and events, their meanings in relation to each other, and their associations with reward and punishment" (Todd & Manaligod 2017, 122–123); this yields dynamically structured models that allow people to figure out how to get around in the world. On the other hand, these models can shape the conceptually articulated thoughts that a person will entertain, as well as the kinds of inferences they are willing to carry out. When these capacities are integrated into a single perspective, they yield typically human ways of picturing the world, which may feel—from the inside—as though they are conceptually organized. Perhaps this is why philosophers have often characterized the human mind in ways that appeal to a Language of Thought, which is organized around "word-sized concepts, sentence-sized intentional states and argument-sized inferences" (Williams 2018, 153). But if my argument is roughly correct, our capacity for conceptually-structured thought is an artifact of our socially situated nature, and the concepts we employ reflect the categories that are salient to the people we interact with. My aim in this section and the sequel is to show that this fact has significant implications for the study of situated forms of social cognition.

4.1 Socially sculpted attention

Let's begin with the growing range of evidence that our understanding of the social world arises through active and ongoing participation in culturally scripted patterns of behavior, which lead to the development of "attention allocation strategies that are consistent with local cultural assumptions" (Park

& Kitayama 2011, 77). In line with the approach I developed in Section 2, this appears to yield priority maps that are shaped by rewards that are received for acting in accordance with local norms, and by criticisms that are received for acting in ways that are socially deviant. Recent data also suggests that we can learn how to think about the nature of social groups by tracking how people interact with one another (Lau et al in press); and here too, there is reason to believe that many of the same reward-driven mechanisms are at play (Klucharev et al 2009, 2011). But no matter how these categories are learned, the actions that people take, and the conversations that they have with one another will shape the structure of the world that they and others inhabit—and this will produce stable patterns in the attitudes and attentional strategies that people acquire as socially situated agents (Kitayama & Uskul 2011, 422). This is just to say that both priority maps and conceptual representations are shaped by ongoing social feedback. For example, as they discuss things that are important, they will tend to converge on shared representations of past events (Hirst et al 2018). And this can even lead to the social suppression of aversive information, and to highlighting positive information; in ways that will produce individual memories that are anchored to the groups that people are part of (Coman & Berry 2015; Coman & Hirst 2012, 2015). These effects on memory, and their effects on priority maps deserve further discussion; and I hope to return to them in a future paper. But for now, I want to turn to the ways in which individual differences in learning histories can yield divergent ways of picturing the social world

Numerous studies have revealed that White, middle class, North Americans tend to converge on attentional strategies that insulate their thinking from contextual factors, and focus their attention on discrete entities; they tend to see individual merit as salient, and contextual factors as background phenomena (Adams et al 2010; Markus & Kitayama 2010; Park & Kitayama 2011). This is not an innate disposition, and it's not a fact about every middle class White American. But it's stable pattern of attunement to the kinds of things that such people typically encounter. Attentional maps that highlight individual achievements are reinforced through practices of praising and blaming individuals for their actions; and they are scaffolded by ongoing engagements with "mobility affording transportation and communication infrastructure, the practice of 'leaving home' in young adulthood, the daily practice of eating from individual place settings, and residence in self contained apartment units" (Adams et al 2010, 283). By routinely experiencing these kinds of concrete social realities and social opportunities, people develop a picture of the world where "exploration, expression, and indulgence of unique, individual feelings" are the primary goods to be pursued (Adams et al 2010, 284). And this typically leads them to "express a desire for mastery, control, achievement, choice, self-expression, or uniqueness" (Markus & Kitayama 2010, 421). These habits and dispositions are ways of picturing a categorically structured world, which is organized by relationships of individual success and failures, and which yields the expectation that self-interested actions will tend to yield such success (Markus & Kitayama 2010, 428).

Converging data from Michael Kraus and his colleagues (2012) suggest that economic and social constraints can also shape an agent's way of picturing the world. They argue that the ongoing experience of economic and social freedom leads to the development of cognitive strategies that are focused on internal states, and that treat these states as the dominant influence on thought and behavior. When people are chronically immersed in environments of relative abundance and elevated social rank, they "are free to pursue the goals and interests they choose for themselves", and they can "pursue these goals and interests relatively free of concerns about their material costs" (Kraus et al 2012, 550). And as a result, middle class and upper class individuals tend to prioritize individualized selves, and assume that behavior is generally caused by individuals, instead of depending on contextual or situational factors. By contrast, where the stability of necessary resources is uncertain and unpredictable, this can lead to the attentional prioritization of contextual and situational factors. So people who live in less stable neighborhoods, who face constant economic instability, and who depend on constantly fluctuating institutional resources often experience the world as socially structured, institutionally constrained, and more limited in social opportunities (Kraus et al 2012, 549). As a result, people who are chronically immersed in these sorts of environments tend to develop attentional strategies that are sensitive to cases of overt social control, and they tend to be aware of the continual recording of their actions in accordance with the preferred frameworks of people in positions of social power.

4.2 Racially sculpted attention

My argument can be summarized as follows. Our habits of attention are shaped by ongoing patterns of feedback, from the people that we interact with, and from our movements through the social world; the ongoing reshaping of our priority maps makes it possible for us to acquire skills that we need to succeed in our local environments; and once we have learned to track all of the relevant social phenomena, our attentional biases can help to minimize the amount of cognitive effort that is required to act in a socially accepted way. This can often be a good thing. But where patterns of exclusion and oppression become entrenched in the material and ideological structures of our cities and social spaces, this same process can yield distorted pictures of the world, which nonetheless *feel like* they represent the world 'as it is'. For example, in contexts where people are presented with racialized imagery in films, novels, and news sources, patterns of attentional salience will begin to stabilize around these aspects of their experience. And this will lead them to orient toward any information that is consistent with this racially structured priority map; but just as importantly, they will tend to suppress information that is at odds with their priority map.

Imagine someone is walking down an unfamiliar alley in an unfamiliar city. A friend has told them that this is a dangerous place to be. And as they look around, they notice familiar markers of threat and danger. Or at least that is how things seem to them. Perhaps they have seen similar things in contexts where they felt scared; or perhaps they merely encountered tales of similar situations in novels, films, or news sources that were saturated with danger and violence. But in any case, they become more aware of their surroundings; they search for lurking threats; and they ruminate on potential encounters with danger. Their muscles grow tense, their heart begins to race, and their rate of respiration increases—all in the service of preparing to manage the expected danger. If this person had never *learned* that this was a potentially dangerous situation, things would probably be quite different. But they possess a robust picture of the world, which is anchored to their learning history, and to past experiences; and this picture affects their threshold for experiencing fear. This might be a good thing, if it helps them avoid a genuine threat to their wellbeing. It might also send them running from a harmless rat that scurries from behind a trash bin. And it may lead them to act on racist or classist biases, causing substantial harm to innocent individuals.

To make this set of claims more concrete, consider the factors that are at play in first-person shooter tasks (FPST), where participants are asked to decide whether someone is holding a gun or an innocuous object (e.g., a cell phone or a wallet), and to use a button press to respond (shoot vs don't shoot). In carrying out this type of task, people must integrate multiple sources of information; some of the relevant information will be conceptual, some will be affectively structured, and some will be organized by habituated expectations (for similar claims in the context of implicit bias, see Amodio 2014; Faucher & Poirer 2017, Van Bavel et al 2012). And the way that these sources of information are integrated will have an impact on the patterns of response that people tend to display in this kind of task.[14] In a meta-analysis of 42 experiments, Yara Mekawi & Konrad Bresin (2015, 124) found that "participants were quicker to shoot armed Black ['suspects], slower to not shoot unarmed Black ['suspects], and were more likely to have a liberal shooting threshold for Black ['suspects]". But while people were more likely to respond by choosing to shoot a Black person (vs a White person), there

[14] Elsewhere, I have defended a computational approach to implicit bias, which explains how these sources of information are likely to be integrated to yield decisions (Huebner 2016, 2018). The view that I defend shares much in common with Edouard Machery's dispositional approach to implicit cognition. Here, I build on important claims that have emerged in the context of recent discussions about the role of attention in cognitive permeation (e.g., Machery 2016 and Gross 2017), and I extend this approach to cover the nearby phenomena known as 'shooter bias', which draws on more robust models (e.g., signal detection theory), and which has received far less philosophical discussion than 'implicit bias'.

were not significant differences in falsealarm rates. This suggests that people are not seeing innocuous objects as guns when they are in the hands of Black people, even though they are more willing to choose to shoot them. To me, this suggests that the problem runs much deeper, and it tells us something significant about the socially sculpted attentional map of race in the US.[15] But in order to see why I believe that this is the case, we will need to look at data showing which kinds of social information are at play in the production of such responses.

To begin with, there is evidence that patterns of response in experiments using FPST are affected by cultural factors (Mekawi & Bresin 2015). For example, people who live in states with more permissive gun laws tend to be more likely to shoot overall; and people who live in cities with lower proportions of White inhabitants tend to display higher levels of anti-Black bias. Contextual factors also seem to matter a great deal (Correll et al 2011): where 'suspects' are situated amid signals of social threat, shooting thresholds are lower across the board, leading to similar patterns of response to white and Black 'suspects'; likewise, when people read stories about white criminals before a FPST, the topography of their attentional landscape shifts, in ways that lead them to respond to both white and Black 'suspects' as equally threatening; finally, a training period where higher numbers of white people are presented with guns, shifts attention toward the assumption that white 'suspects' are more likely to have a gun. Presumably, the second and third effects are short term artifacts of the experimental environment. But the first effect is likely to depend on a robust positive association between proportion of Black people in a neighborhood and white people's fear of crime (Chiricos, Hogan, & Gertz, 1997). The presence of a high number of Black people in a neighborhood is often sufficient to trigger the assumption that a neighborhood is dirty, disordered, and dangerous (Sampson & Raudenbush 2004). And where people feel like a situation or person is dangerous, their threshold for responding to a potential threat will be greatly reduced.

Even more strikingly, Joshua Correll and his colleagues (2015) have shown that attention is likely to play a critical role in FPSTs. Using an eye tracker, they found a significant effect of race on the visual angle between the object to be categorized (the weapon or the innocuous object) and the fovea at the time of response. More specifically, the visual angle was larger at the point where a decision was made about whether to shoot a Black person (Black,

[15] In a study that monitored event-related potentials (ERPs) during a FPST task, Correll et al (2006) found that a larger P200 response to images of Black people predicted the extent to which people were quicker to choose to "shoot" armed Black 'suspects', and to "not shoot" unarmed white 'suspects'. Similarly, Amodio et al (2004) found larger event-related negativity (which they interpret as activity in the anterior cinculate cortex) where race and object are stereotypically incongruent, suggesting perceived conflict; moreover, they found that more pronounced event related negativity (ERN) correlated with greater accuracy and slower reaction times, suggesting that increased cognitive control could be used to inhibit this prepotent response.

M=2.08°; White, M=1.59°); and this difference persisted in contexts where a weapon was present (Black, M=2.03°; White, M=1.41°). This suggests that people need more precise visual information to decide whether to shoot a white person, and more precise information to see whether they are holding a gun. This is what we should expect if most people possess a priority map that highlights Black people as threatening. Put much too simply, people are more likely to choose to shoot a Black person because their attentional landscape highlights the connection between Black people and danger; so they need less visual information to decide in favor of the hypothesis that a Black person is dangerous because they possess a picture of the world that highlights the connection between race and threat (cf., Machery 2016, 64).

Intriguingly, things look different when police officers take part in FPSTs. They tend to be quicker, more accurate, and more sensitive to the presence of guns (Correll 2007). This makes sense, as they are typically trained to hold their fire when they are uncertain; and they often cultivate forms of reflexive control, using training conditions where people fire paintballs or simulated and painful ammo. And as a result, they should be more likely to ignore irrelevant information, and to focus on situationally relevant stimuli. Even so, police officers tend to be slower to respond to counter-stereotype situations; and those who work in communities where there are larger Black populations, and higher levels of violent crime, tend to have more biased response latencies. And crucially where their work environment reinforces the salience of the connection between race and violence, high levels of training are insufficient to mitigate the effects of racial bias. Like untrained community members, officers who work in Gang Units and Violent Crime units are more likely to choose to shoot Black than White 'suspects' in a first person shooter task (Sim et al 2013, 300).

Of course, these kinds of tasks are not ecologically valid, and it is difficult to know what to infer from FPST tasks. Indeed, a recent experiment using a more realistic and more immersive simulation, where people had to decide whether to shoot a laser-equipped handgun, found that people were more likely to 'shoot' unarmed white 'suspects' than unarmed Black 'suspects (46/184 vs 1/47; James et al 2014). These data seem to contradict the data that have been collected on computers using FPST tasks. But in thinking about this experiment, is important to note that this was a task that required a decision to shoot or not, whereas standard FPSTs require a decision about whether to push a button on the left or the right. And critically, participants in this task took longer to decide whether or not to shoot Black 'suspects' (even when they were armed); and data collected using electroencephalogram (EEG) during the simulation revealed higher levels of alpha-wave suppression for armed as well as unarmed Black 'suspects'. There are multiple ways of interpreting these data, but the most plausible hypothesis is that Black 'suspects' were always perceived as threatening, that the alpha-wave suppression reveals an inhibitory response, and that the slower response reveals active suppression of fear as a result of the desire not to appear racist. This strikes me as plausible, because the goal of not appearing racist is likely to play a much more prominent role in shaping a decision about whether to act or refrain from acting.

It would require a more substantial argument to establish this claim conclusively. But so far as I can tell, it accords with all of the existing data, as well as the argument I have developed throughout this paper. More importantly, it suggests that if we could always trust people to inhibit their initial responses, it would be possible to reshape problematic patterns of socially entrenched behavior merely by cultivating these kinds of goals. Unfortunately, I'm not optimistic that these data provide insights into human behavior outside of laboratory environments. To prevent the emergence of racial bias, people would need to cultivate highly salient and conceptually structured goals of inhibiting racially charged responses, and this goal would need to play an ongoing role in real life decisions. I doubt that this is likely to be the dominant motivation in real-life situations for most people; and in the absence of strong anti-racist motivations, and active attempts to suppress unexpected and problematic responses, we have little reason to believe that these kinds of attentional effects will generalize to the kinds of situations that we should be concerned about (cf., Huebner 2016).

4.3 The production of racially sculpted concepts

With this background in hand, I want to turn in closing to a suggestion about how racialized concepts become stable in contexts where they are experienced as highly salient. To begin with, note that people in the United States tend to overestimate the percentage of the population who are Black, Jewish, Asian, and Latinx; people from communities with larger white populations tend to guess that the nation has a larger white population; and people from communities with larger Black populations tend to guess that the nation has a larger Black population (Wong 2007, 401). But shifts in the evaluative salience of different groups can complicate this situation in troubling ways. As Charles Gallagher (2003) argues, people who live in communities where white people interact primarily with other white people (which is the norm in the US) will often acquire highly distorted pictures of the national population. He focuses on three situations that yield the salient distortions.

1. Some people who live in predominantly white spaces encounter racialized minorities primarily in the context of films and news media that present the Black population as dangerous; this heightens their experience of racial anxiety, increasing the salience of both real and virtual encounters with Black people, and causing something like oversampling in perception and memory. Strikingly, Gallagher's data show that this can lead to extreme over-estimations of the Black population (40–60%; actual 12%).

2. Other people who encounter racialized minorities find collective demands for racial justice to be more salient. This can focus attention on presentations of Black activists in the news media, and on social situations where demands for change are being made; and here too, people tend to substantially overestimate the percentage of the US population that is Black. And just as importantly, it can lead them to make illicit assumptions about Black people being better off than white people in achieving their needs and interests.
3. Finally, in contexts where people become anxious about demographic shifts that could make the US a majority-minority nation, this can heighten the salience of encounters with Black people, and again lead to a situation where people will tend to overestimate the proportion of the population that is Black.

These specific effects are distinctive of the US, where a specific history of racial injustice makes race a highly salient feature of the social environment. However, similar effects are likely to emerge in any context where there are interactions between statistical stabilities and socially reinforced evaluations, and wherever the salience of a group recruits attention, shapes memories, and affects the decisions that people make. Given the coalitional nature of human psychology, this means that similar kinds of phenomena are likely to show up across most populations (Van Bavel & Pereira 2018). One place where this affect is obvious is on the kinds of assumptions and inferences that are beginning to emerge and solidify around issues of immigration, in many parts of the world. Paralleling my discussion about 'coffee' above, there are multiple routes to the acquisition and use of a socially salient term like 'immigrant' and the meaning of these terms can't be specified by appeal to any of these causal relations. Terms like 'immigrant', along with countless other terms, mean what they do in virtue of their use in practices of functional categorization; but social forces can lead to the emergence of highly divergent ways of using such a term, which often become clear only after sustained discussion, which often becomes highly unpleasant. For some people, 'immigrant' may be little more than a linguistic label, which is anchored to things they've heard on talk radio, seen in the news, or read on a presidential twitter feed; or it might derive from the inflammatory rhetoric of a political campaign. Such people will become more attentive to information that accords with this acquired picture of the world, and their thoughts about immigration will be driven by anxiety. For other people, the inferential structure of 'immigrant' may be more highly elaborated, as it may be shaped by experiences with friends, colleagues, or random people throughout the city; they will have a richer understanding of why people move to a new country, as well as a more robust understanding of the roles that immigrants play in a vibrant and thriving society. These kinds of differences can produce highly divergent ways of using a term like 'immigrant'. The fact that we use the same word, however, can often lead us to believe that we picture the world in the same way. And this can lead us to neglect the causal and structural forces that sustain habits of thought and attention.

5. Concluding thoughts

If the argument that I have developed is roughly right, agents learn to situate themselves within the world *they encounter* (cf., Sellars 1978 §28–29). And their ways of picturing the world shape the storage and retrieval of memories, leading to patterns of thought and behavior that are socially shaped and sustained. But local patterns of attunement can yield global patterns of distortion. And where they do:

> We compare, struggle, and wonder how to let go of our personal, subjective view and arrive at an objective recognition of things. We want to be directly in touch with the reality of the world. Yet the objective reality we think exists independently of our sense perceptions is itself a creation of collective consciousness. Our ideas of happiness and suffering, beauty and ugliness, are reflections of the ideas of many people (Thích Nhất Hạnh 2006, 39).

This is not something that is unique to specific ways of encountering the world: all typically functioning humans will develop strategies for prioritizing information, in accordance with their long and short-term goals (Todd & Manaligod 2017, 122). And given the socially structured nature of our goals, they will all tend to operate from within their own pictures of the world. Even so, the fact that these ways of thinking are constructed means that it's possible to reshape what is salient, and to reorient our habituated tendencies to act in specific ways (see Cikara and Van Bavel 2014 for a review). Even the most socially-entrenched biases become less pronounced in contexts where an alternative way of categorizing is available. For example, where people attend to the shared love of a football team, or to shared commitments to a university, their attentional biases shift toward these categories (Van Bavel & Cunningham, 2009), at least so long as these categories are the contextually salient way of dividing up the world. When we look at the world from a different perspective, the features that are most salient will begin to shift. The key question is: what would it take to make such a shift permanent?

Unfortunately, I'm not sure that this is possible. Though there is plenty of room for empirical research on ways of re-shaping these kinds of habituated attentional biases. I believe that the most promising option will require acknowledging the deep respects in which all human interests are interconnected and interdependent. This point is defended in some parts of Buddhist philosophy; and it is nicely summed up in an Ashanti metaphor about a crocodile with two heads, which are fighting with one another over access to food (Wiredu 1995, 57). If the two heads were to recognize that any food that either of them ate would end up in the same stomach, then the motivation to compete would dissipate, and it would be replaced by cooperative drives for mutual aid and mutual support. According to Ashanti tradition, human conflict can always be reconciled through patterns of dialogue, which highlight shared needs and shared interests. So long as everyone listens, and so long as they all work to build a shared understanding of the situation, the drive toward mutual aid and mutual support will always arise. Of course, dialogue across deep cultural divides is never easy. And in many cases, we get caught up in attempting to defend our own beliefs, without listening to the things that other people need. This occurs both in the context of interpersonal relations, and in the context of cultural differences. When we become anxious, it is harder to be vulnerable, and it is harder to see points where new paths forward can be developed. There is evidence that feelings of racial anxiety can trigger an increase in the release of norepinephrine, which can compromise cognitive control, and lead to forms of thought and behavior that are more directly shaped by the implicit structure of priority maps (cf., Amodio et al 2004; Godsil & Richardson 2016). As I noted above, differences in the availability of norepinephrine can shift the affective salience of different kinds of stimuli, and they can shape what kinds of features of the world we attend to. So we need to find ways of mitigating anxiety; and this will require either contemplative training, or prefigurative forms of social practice (and maybe both). But that's another story for another day.

6. Works cited:

Abraham, T. (2019). Cybernetics. In *The Routledge Handbook of the Computational Mind*. M. Sprevak & M. Colombo, eds. Routledge.

Adams, G., Salter, P. S., Pickett, K. M., Kurtis, T., & Phillips, N. L. (2010). Behavior as mind in context. *The mind in context*, 277–306.

Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences* 1191, 42–61.

Akins, K. (1996). Of sensory systems and the "aboutness" of mental states. *The Journal of Philosophy*, *93*(7), 337–372.

Amodio, D. M. (2014). Dual Experiences, Multiple Processes: Looking Beyond Dualities for Mechanisms of the Mind. In J. S. Sherman, B. Gawronski & Y. Trope (eds.), *Dual Process Theories of the Social Mind*, NY: Guilford Press, 560–576.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science*, 15(2), 88–93.

Anderson, B. (2016). The attention habit: How reward learning shapes attentional selection. *Annals of New York Academy Sciences*, 1369 (1), 24–39.

Awh, E., Belopolsky, A.V. &Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, 16 (8), 437–443.

Chiricos, T., Hogan, M., & Gertz, M. (1997). Racial composition of neighborhood and fear of crime. *Criminology*, *35*(1), 107–132.

Chomsky, N. (1995). Language and nature. *Mind*, *104*(413), 1–61.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: an integrative review. *Perspectives on Psychological Science*, 9(3), 245–274.

Clark, A. (1996). Linguistic anchors in the sea of thought?. *Pragmatics & Cognition*, 4(1), 93–103.

Clark, A. (2014). *Mindware*. Second Edition. Oxford: Oxford University Press.

Coman, A., & Berry, J. N. (2015). Infectious Cognition: Risk perception affects socially shared retrieval-induced forgetting of medical information. *Psychological Science*, 26(12), 1965–1971.

Coman, A., & Hirst, W. (2012). Cognition through a social network. The propagation of induced forgetting and practice effects. *Journal of Experimental Psychology: General*, 141(2), 321–33.

Coman, A., & Hirst, W. (2015). Social identity and socially shared retrieval-induced forgetting: The effects of group membership. *Journal of Experimental Psychology: General*, 144(4), 717–722.

Correll, J., Urland, G. R., & Ito, T. A. (2006). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology*, *42*(1), 120–128.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: police officers and racial bias in the decision to shoot. *Journal of personality and social psychology*, *92*(6), 1006.

Correll, J., Wittenbrink, B., Park, B., Judd, C. M., & Goyle, A. (2011). Dangerous enough: Moderating racial bias with contextual threat cues. *Journal of experimental social psychology*, *47*(1), 184–189.

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of personality and social psychology*, 108(2), 219.

Craik, K. (1967). *The nature of explanation. 1943*. Cambridge University, Cambridge UK.

Crockett, M. (2013). Models of morality. *Trends in Cognitive Science*, 17, 8, 363–6.

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, 25(5), 736–760.

Cushman, F. (2013). Action, outcome and value: A dual-system framework for morality. *Personality and Social Psychology Review*, 17 (3), 273–292.

Davidson, D. (1986). A nice derangement of epitaphs. *Philosophical grounds of rationality: Intentions, categories, ends*, *4*, 157–174.

De Brigard, F., Brady, T. F., Ruzic, L., & Schacter, D. L. (2017). Tracking the emergence of memories: A category-learning paradigm to explore schema-driven recognition. *Memory & cognition*, 45(1), 105–120.

De Brigard, F., Hanna, E., St Jacques, P. L., & Schacter, D. L. (2018). How thinking about what could have been affects how we feel about what was. *Cognition and Emotion*, 1–14.

Delgado, M. R., Nearing, K. I., LeDoux, J. E., & Phelps, E. A. (2008). Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. *Neuron*, *59*(5), 829–838.

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in human neuroscience*, *11*, 481.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, *170*(1), 115–135.

Egan, F. (2019). The nature and function of content in computational models. In *The Routledge Handbook of the Computational Mind*. M. Sprevak & M. Colombo, eds. Routledge.

Faucher, L. & Poirer, P. (2017). Mother culture, meet mother nature. In Huebner, B. (Ed.). (2017). *The Philosophy of Daniel Dennett*. Oxford University Press.

Gallagher, C. A. (2003). Miscounting race: Explaining Whites' misperceptions of racial group size. *Sociological Perspectives*, *46*(3), 381–396.

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (submitted). Rethinking dopamine prediction errors.

Gershman, S. J. (in press). Uncertainty and exploration. *Decision*.

Godsil, R. D., & Richardson, L. S. (2016). Racial Anxiety. *Iowa L. Rev.*, *102*, 2235.

Gross, S. (2017). Cognitive penetration and attention. *Frontiers in psychology*, *8*, 221.

Haas, J. (2018). An empirical solution to the puzzle of weakness of will. *Synthese*, 1–21.

Hale, K. (1986) Notes on World View and Semantic Categories: Some Warlpiri Examples, in *Features and Projections*. P. Muyskens & H. van Riemsdijk (eds). Dordrecht: Foris, 233–54.

Hebb, D. O. (1949). *The organization of behavior: A neurophysiological approach*. Wiley.

Henderson, R. & McCready, E. (in press). How dogwhistles work. *The proceedings of LENLS*.

Hirst, W., Yamashiro, J. K., & Coman, A. (2018). Collective Memory from a Psychological Perspective. *Trends in Cognitive Sciences*, 22(5), 438–451.

Huebner, B. (2016). Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. In Brownstein, M. & J. Saul (Eds.). *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*. Oxford University Press, 47–79.

Huebner, B. (2018). Reply to Del Pinal and Spaulding. In a symposium on "Conceptual Centrality and Implicit Bias" at *The Brains Blog*. https://goo.gl/2eV3je

Huebner, B. (in press). The interdependence and emptiness of whiteness. In *Buddhism and whiteness*. E. McRae & G. Yancy, eds. Lexington Books.

James, L., Klinger, D., & Vila, B. (2014). Racial and ethnic bias in decisions to shoot seen through a stronger lens: Experimental results from high-fidelity laboratory simulations. *Journal of Experimental Criminology*, 10(3), 323–340.

Jiang, Y. V. (2018). Habitual versus goal-driven attention. *Cortex*, *102*, 107–120.

Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., White, J. P., Ellis, T. L., Phillips, P. E. and Montague, P. R. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences*, 113(1), 200–205.

Kitayama, S., & Uskul, A. K. (2011). Culture, mind, and the brain: Current evidence and future directions. *Annual review of psychology*, 62, 419–449.

Klucharev V., Hytönen K., Rijpkema M., Smidts A., Fernández G. (2009). Reinforcement learning signal predicts social conformity. *Neuron* 61, 140–151

Klucharev V., Munneke M., Smidts A., Fernández G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. *J. Neurosci*. 31, 11934–1194

Kraus, M. W., Piff, P. K., Mendoza-Denton, R., Rheinschmidt, M. L., & Keltner, D. (2012). Social class, solipsism, and contextualism: how the rich are different from the poor. *Psychological review*, 119(3), 54.

Krauzlis, R. J., Bollimunta, A., Arcizet, F., & Wang, L. (2014). Attention as an effect not a cause. *Trends in cognitive sciences*, 18(9), 457–464.

Krosch, A. R., & Amodio, D. M. (2014). Economic scarcity alters the perception of race. *Proceedings of the National Academy of Sciences*, *111*(25), 9079–9084.

Kryklywy, J. H., & Todd, R. M. (2018). Experiential History as a Tuning Parameter for Attention. *Journal of Cognition*, 1(1), 24.

Kukla, R. (2017). Embodied Stances: Realism without Literalism. In *The Philosophy of Daniel Dennett*. B. Huebner, ed. New York: Oxford University Press, 3–31.

Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (in press). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General*.

Lee, D., Todd, R. M., Gardhouse, K., Levine, B., & Anderson, A. K. (2013). Enhanced attentional capture in survivors of a single traumatic event. In *Society for Neuroscience Annual Meeting, San Diego, CA, USA*.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, *47*(11), 1940–1951.

Levine, S. M. (2007, September). The place of picturing in Sellars' synoptic vision. In *The Philosophical Forum* (Vol. 38, No. 3, pp. 247–269). Malden, USA: Blackwell Publishing Inc.

Lightfoot, D. (2017). Invariant and variable properties. Inference, 3, 2. Retrieved from http://inference-review.com/article/invariant-and-variable-properties on 31 August 2018.

Lindquist, K. A. (2013). Emotions emerge from more basic psychological ingredients: A modern psychological constructionist model. *Emotion Review*, 5(4), 356–368.

Machery, E. (2016). Cognitive penetrability: a no-progress report. Zeimbekis, J., & Raftopoulos, A. (Eds), *The cognitive penetrability of perception. New philosophical perspectives*. New York: OUP.

Markus, H. R., & Kitayama, S. (2010). Cultures and selves: A cycle of mutual constitution. *Perspectives on Psychological Science*, 5(4), 420–430.

Maturana, H. R., Lettvin, J. Y., McCulloch, W. S., & Pitts, W. H. (1959). Evidence that cut optic nerve fibers in a frog regenerate to their proper places in the tectum. *Science*, *130*(3390), 1709–1710.

McCready, E. (2012) Emotive equilibria. *Linguistics and Philosophy* 35, 243–283.

Mekawi, Y., & Bresin, K. (2015). Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology*, 61, 120–130.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan P. (2012). Computational psychiatry. *Cognitive Science* 16: 72–80.

Nhất Hạnh, T. (2006). *Understanding our mind*. Parallax press.

Newell, A., & Simon, H. (1956). The logic theory machine--A complex information processing system. *IRE Transactions on information theory*, *2*(3), 61–79.

Newell, A. & Simon, H. (1976). Computer Science as Empirical Inquiry: Symbols and Search, *Communications of the ACM*, 19 (3), 113–126

O'Shea, J. (2007). *Wilfrid Sellars: Naturalism with a normative turn*. Polity, Cambridge.

Olen, P. (2018). The Varieties and Origins of Wilfrid Sellars' Behaviorism. In *Sellars and the History of Modern Philosophy*. L. Corti & A. Nunziante, eds. Routledge.

Park, H., & Kitayama, S. (2011). Perceiving through culture: The socialized attention hypothesis. In N. Ambady, K. Nakayama, S. Shimojo and R. B. Adams, Jr. (Eds.), *Social Vision*. New York: Oxford University Press.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. OUP Oxford.

Piccinini, G. & Bahar, S. (2013). Neural Computation and the Computational Theory of Cognition, *Cognitive Science*, 37 (3), 453–488.

Piccinini, G. & Scarantino, A. (2011). Information Processing, Computation, and Cognition, *Journal of Biological Physics*, 37 (1), 1–38.

Thomson, E., & Piccinini, G. (2018). Neural Representations Observed. *Minds and Machines*, *28*(1), 191–235.

Poser, W. J. (1992). Blocking of phrasal constructions by lexical items. *Lexical matters*, 111–130.

Railton, P. (2014). Reliance, Trust, and Belief. *Inquiry*, 57(1), 122–150.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Sachs, C. B. (2018). In Defense of Picturing: Sellars's Philosophy of Mind and Cognitive Neuroscience. https://doi.org/10.1007/s11097–018–9598–3

Salisbury, J. & Schneider, S. (2019). Concepts, symbols and computation: An integrative approach. In *The Routledge Handbook of the Computational Mind*. M. Sprevak & M. Colombo, eds. Routledge.

Sampson, R. J., & Raudenbush, S. W. (2004). Seeing disorder: Neighborhood stigma and the social construction of "broken windows". *Social psychology quarterly*, *67*(4), 319–342.

Schneider, S. (2009). The Nature of Symbols in the Language of Thought, *Mind and Language*, 24, 4, 523–553.

Schonbein, W. (2012). The Linguistic Subversion of Mental Representation. *Minds and Machines*, *22*(3), 235–262.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80 (1), 1–27.

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behavioral and brain functions*, 6(1), 1.

Selfridge, O. (1958). Pandemonium: a paradigm for learning. In *Symposium on the Mechanization of thought Processes,* London: HM Stationery Office.

Sellars, W. (1956). Empiricism and the Philosophy of Mind, in *Minnesota Studies in The Philosophy of Science*, Vol. I, H. Feigl & M. Scriven, eds. Minneapolis: University of Minnesota Press, 253–329.

Sellars, W. (1960). Being and Being Known, *Proceedings of the American Catholic Philosophical Association*, 28–49.

Sellars, W. (1974). Meaning as Functional Classification. *Synthese*, 27, 417–37

Sellars, W. (1978). The Role of Imagination in Kant's Theory of Experience. In *Categories*. H. Johnstone, Jr. (ed.). Pennsylvania State University, 231–45

Sellars, W. (1981). Mental Events. *Philosophical Studies* 39, 325–45.

Shah, A. K., Zhao, J., Mullainathan, S., & Shafir, E. (2018). Money in the mental lives of the poor. *Social Cognition*, 36(1), 4–19.

Sim, J. J., Correll, J., & Sadler, M. S. (2013). Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot. *Personality and social psychology bulletin*, 39(3), 291–304.

Spencer, M. K. (2015). The Category of Habitus: Accidents, Artifacts, and Human Nature. *The Thomist: A Speculative Quarterly Review*, 79(1), 113–154.

Talmi, D., Ziegler, M., Hawksworth, J., Lalani, S., Herman, C. P., & Moscovitch, M. (2013). Emotional stimuli exert parallel effects on attention and memory. *Cognition & emotion*, 27(3), 530–538.

Theeuwes, J. (2018). Visual Selection: Usually Fast and Automatic; Seldom Slow and Volitional. *Journal of Cognition*, 1(1), 21–29.

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

Todd, R. M., & Manaligod, M. G. (2017). Implicit guidance of attention: The priority state space framework. *Cortex*, 30(1), e1–8.

Todd, R. M., Müller, D. J., Palombo, D. J., Robertson, A., Eaton, T., Freeman, N., Levine, B., Anderson, A. K. (2013). Deletion variant in the ADRA2B gene increases coupling between emotional responses at encoding and later retrieval of emotional memories. *Neurobiology of Learning and Memory*, 112, 222–229.

Todd, R. M., Cunningham, W. A, Anderson, A. K., & Thompson, E. (2012). Affect-biased attention as emotion regulation. *Trends in Cognitive Sciences*, 16(7), 365–72.

Todd, R. M., Ehlers, M. R., Mueller, D. J., Robertson, A., Freeman, N., Palombo, D. J., Levine, B., & Anderson, A. K. (2015a). Neurogenetic variations in norepinephrine availability enhance perceptual vividness. *Journal of Neuroscience 35* (16), 6506–6516.

Todd, R. M., MacDonald, M. J., Sedge, P., Robertson, A., Jetly, R., Taylor, M. J., & Pang, E. W. (2015). Soldiers with posttraumatic stress disorder see a world full of threat: magnetoencephalography reveals enhanced tuning to combat-related cues. *Biological psychiatry*, *78*(12), 821–829.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. University of California Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433.

Van Bavel, J. J., & Cunningham, W.A. (2009). Self-categorization with a novel mixed-race group moderates automatic social and racial biases. *Personality and Social Psychology Bulletin*, 35(3), 321–335.

Van Bavel, J. J., J. Xiao & W. A. Cunningham. (2012). Evaluation is a Dynamic Process: Moving Beyond Dual System Models. *Social and Personality Psychology Compass*, 6/6, 438–454.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An Identity-based model of political belief. *Trends in cognitive sciences*.

Whitman, J. C., Zhao, J., Roberts, K. H., & Todd, R. M. (2018). Political orientation and climate concern shape visual attention to climate change. Climatic Change, 147(3–4), 383–394.

Wiener, N. (1948). *Cybernetics: Control and communication in the animal and the machine*. Wiley.

Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines*, *28*(1), 141–172.

Wong, C. J. (2007). “Little” and “big” pictures in our heads: Race, local context, and innumeracy about racial groups in the United States. *Public Opinion Quarterly*, *71*(3), 392–412.

Zhao, J., Al-Aidroos, N., & Turk-Browne, N. B. (2013). Attention is spontaneously biased toward regularities. *Psychological Science*, 24(5), 667–677.

Carrie Figdor
Department of Philosophy
University of Iowa

Original Scientific Paper
UDK 165.194:159.954/.956
165.3/.6

THE FALLACY OF THE HOMUNCULAR FALLACY[1]

Abstract. *A leading theoretical framework for naturalistic explanation of mind holds that we explain the mind by positing progressively "stupider" capacities ("homunculi") until the mind is "discharged" by means of capacities that are not intelligent at all. The so-called homuncular fallacy involves violating this procedure by positing the same capacities at subpersonal levels. I argue that the homuncular fallacy is not a fallacy, and that modern-day homunculi are idle posits. I propose an alternative view of what naturalism requires that reflects how the cognitive sciences are actually integrating mind and matter.*

Keywords: *homuncular functionalism, homuncular fallacy, mechanistic explanation, psychological models, naturalizing the mind, psychological explanation*

1 Introduction

The most familiar, and arguably received, theoretical framework for an adequate naturalistic explanation of mind is homuncular functionalism (Lycan 1991; Dennett 1975; Fodor 1986; Cummins 1983).[2] The homuncular functionalist strategy is to explain a cognitive capacity of a whole in terms of the organized not-quite-cognitive operations of its parts. Because of its part-whole decompositional nature, homuncular functionalism can be seen as a special case of mechanistic explanation, a leading contemporary view of scientific explanation (e.g., Machamer et al. 2000; Bechtel 2005; Glennan 2002; Craver and Tabery 2015).[3] Unfortunately, homuncular functionalism cannot

[1] This article builds on a position initially articulated and defended in Ch. 8 ("Literalism and Mechanistic Explanation") of Figdor (2018). The book offers a comprehensive discussion of the problem of interpreting psychological language throughout biology.

[2] Throughout, I will use "cognitive", "psychological" and "mental" interchangeably, and "capacities", "functions", "operations", and "activities" interchangeably to denote what entities are able to do and what they actually do on occasion. Nothing here turns on the difference between ascribing to an entity an ability to do X or to perform a function X and describing an entity as doing X or performing the function X.

[3] I will discuss homuncular functionalism's relation to ongoing debates in psychological explanation in more detail in section 3. In brief, however, homuncular functionalism counts as a type of functional analysis, but its assignment of the homunculi to components makes the framework mechanistic (Craver 2001; Piccinini and Craver 2011)

doi: 10.5937/BPA1831041F

bask in the popularity of the new mechanistic philosophy. Contemporary mechanism, alongside other developments, shows that modern-day homunculi are idle posits, just as their Aristotelian namesakes once were to historical mechanists. Even further, the mind is being naturalized using explanatory tools that do not require them.

In what follows I first present the mechanistic explanatory framework defended in recent years and its relation to homuncular functionalism. I then show how accepted mechanistic explanations regularly violate the characteristic constraints of homuncular functionalism without triggering the epistemic disaster that motivates the framework. In addition, there is no explanatory justification for a psychological exception to the rule. As a result, homunculi are bogeymen in principle and are treated as such in scientific practice. I conclude by suggesting an alternative view of what naturalism requires in the context of contemporary efforts to integrate psychology and neuroscience.

2 Homuncular Functionalism and Mechanistic Explanation

The term "homunculus" was originally introduced to label the preformed little men posited within the Aristotelean explanatory tradition to explain how the human adult could emerge from an embryo (Maienschein 2017). Homunculi resolved the problem by holding that everything in the adult is already in the embryo – not in the sense accepted today that the adult *develops* from what is already in the embryo (ignoring the role of environmental factors), but in the sense that the embryo already contains the final result in miniature. A kernel of this view persists today in the claim that an embryo (or even a zygote) is an unborn child with rights.

Seventeenth-century supporters of the then-new mechanistic approach to explaining nature rejected Aristotelian homunculi as idle posits. In general terms, both historically and contemporaneously, a mechanistic explanation is an explanation of a capacity of an entity in terms of that entity's constituent entities and activities and their organized causal interactions (e.g., Machamer, Darden, and Craver 2000, Glennan 2002, Bechtel and Abrahamsen 2005). Moliere's ridiculing reference in *Le Malade Imaginaire* to the dormitive virtue of a sleeping potion was of a piece with this rejection of homunculi by mechanists. Fodor (1968: 627) recapitulates this ridicule for contemporary readers with a farcical computational account of how to tie one's shoes in terms of a little man inside who follows a rulebook for tying one's shoes.

Historical mechanists also understood mechanisms in a narrow way that reflected the machines with which they were familiar, such as clocks or

even if psychological explanation in general may not be (Weiskopf 2011). According to functionalism in philosophy of mind, which I assume here for the sake of argument, mental states are individuated by their functional roles in a cognitive system.

mills. This narrow conception informed La Mettrie's (1748/2017) defense of mechanistic explanation of mind, as well as Leibniz's rejection of it. To Leibniz (1714/1989), "perception, and what depends on it, is inexplicable in terms of mechanical reasons, that is, through shapes and motions": one could imagine walking into enlarged version of a machine structured to make it think and perceive, and inside "we would only find parts that push one another, and we will never find anything to explain a perception." Homuncular functionalists side with La Mettrie in terms of the possibility of a mechanistic explanation of mind, but the lingering question has been how to bridge the explanatory gap that Leibniz and others identified.

To do this, homuncular functionalism imposes three characteristic adequacy conditions on its decompositional explanations of cognitive capacities. First and second, the subcapacities ascribed to the parts must be distinct from and simpler than the cognitive capacity of the whole that they explain. Lycan explains both conditions as follows:

> [H]omunculi can after all be useful posits, so long as their appointed functions do not simply parrot the intelligent capacities being explained. ... We account for the subject's intelligent activity, not by idly positing a single homunculus within that subject whose job it simply is to perform that activity, but by reference to a collaborative team of homunculi, whose members are individually more specialized and less talented. (Lycan 1991: 260)

The homunculi are also variously described as less problematic, elementary, primitive, or less clever (Cummins 1983; Dennett op.cit.). These two features give rise to their being called homunculi, although the label is a bit misleading: the original homunculi were exactly similar in all relevant respects to what they were miniatures of, whereas modern-day homunculi are explicitly barred from being exactly similar in these respects.

Third, the call for less demanding subfunctions must be iterated at each level of the decompositional explanation until the subfunctions are no longer cognitive and the homunculi are "discharged" (Dennett 1975). As Fodor (1968: 629) puts it, a modern-day homunculus is a "representative *pro tem*" for a system of instructions "that makes no reference to unanalyzed psychological processes". The iteration of subfunctions continues until the activities of the smallest parts within the explanation's scope are entirely devoid of intelligence. Further decompositional explanation – to the levels of cellular mechanisms, for example – proceeds by means of capacities that are entirely non-cognitive (i.e., physical). This gradualism enables modern-day homunculi escape Molierian/Fodorian ridicule; it is the cognitive analogue of the embryo-to-adult relationship as now understood. Since we already accept a naturalistic theory of how an embryo can develop into an adult, it seems but a small step to understand a framework that posits a similar relation between physical and psychological capacities, at least synchronically (and maybe diachronically).

These three adequacy conditions are motivated by the epistemic worry that ascribing the psychological capacity of a whole to a part is non-explanatory or idle, as the uselessness of the original homunculi in a mechanistic framework appeared to demonstrate. Bechtel (2009: 561) writes that "assuming a homunculus with the same capacities as the agent in which it is posited to reside clearly produces no explanatory gain", while Dennett (1975: 171) adds that to avoid being "question-begging" the most fundamental posits within the explanation must not be supposed to perform tasks or follow procedures requiring intelligence.

The homuncular fallacy involves violating these requirements for distinct, ever-stupider capacities in a decompositional series that gradually gives way to non-cognitive capacities. To commit the fallacy is to fail to naturalize the mind.

3 Troubles for Homuncular Functionalism

The homuncular functionalist's adequacy conditions have not gone unchallenged. For example, Margolis (1980) argues that there is no *a priori* reason to restrict explaining capacities to "stupider" ones. Freudian psychology posited a complex unconscious which explained simple or complex patterns of behavior. Although the Freudian explanatory framework may have been rejected by many, it was not rejected because it violated the second adequacy condition.

I will challenge homuncular functionalism instead by undermining the epistemic worry that motivates positing homunculi in the first place. The contemporary explanatory practices that motivate the new mechanism also show that the adequacy conditions of homuncular functionalism are unnecessary.

First, contemporary mechanists do not restrict mechanistic explanation to "exclusively mechanical (push-pull) systems" (Machamer, Darden, and Craver 2000: 2). For example, restriction to push-pull mechanisms "makes the concept of a mechanism too narrow to accommodate the diverse kinds of mechanism in contemporary neuroscience" (Craver 2007: 4). Because their framework is intended to capture actual scientific explanatory practices, contemporary mechanists emphasize a characteristic explanatory method rather than characteristic kinds of activities. Margolis' criticism of homuncular functionalism is, in effect, a special case of this point.

Second, these same explanatory practices regularly involve ascribing the same capacities to wholes and their parts without generating epistemic distress. The same kinds of activities appear at different spatiotemporal scales and different decompositional levels, and their contributions to these explanations are not idle. A piece of machinery lifts boxes because its engine has a camshaft that lifts valve covers. If a piano string is vibrating (exhibiting simple harmonic motion), this will be in part because the molecules in the string also vibrate. A monkey learns to swing through the trees by grabbing and releasing vines, and presynaptic neurons in the monkey's hippocampus release glutamate in the process of long-term potentiation, theorized to be a mechanism of learning. Sober (1982:421) raises this point when he notes that the apparent "emptiness" of homuncular explanations does not stem from ascriptions of the same operations to parts and wholes, but from a shift from tokens to types. To borrow his example, no epistemic error is generated when we hold that planets rotate in part because the nuclei of their atoms rotate. In a similar vein, Cartwright (1983: 145) notes the repeated use of the harmonic oscillator model at multiple scales. To the extent that such repetition is problematic, it is because we are no closer to an explanation of (e.g.) rotating as a type of activity ("what unites planets and atoms"). But that's not a reason to deny that planets and their atomic nuclei both rotate.

It follows from these two points that cognitive capacities are just as apt as any other type of capacity for being ascribed to both wholes and their parts, at least in principle.

Third, there is actual disregard of the first and second homuncular functionalist restrictions in scientific psychological practice. For example, Sutton and Barto (1981) theorized about adaptive elements that would comprise adaptive systems, intending their temporal difference (TD) model of reinforcement learning to apply to both. The model ascribes such capacities as anticipating and predicting to the adaptive systems that exhibit this form of learning. Although the model was developed based on animal learning data (including humans), they explicitly suggest neural assemblies as possible targets of their model in addition to humans, dogs, and other adaptive systems with which we are more familiar. Two decades later, Suri and Schultz (2001) implemented the TD model in a connectionist simulation of actual neural activation patterns, showing that Sutton and Barto's speculative suggestion was not idle. Similarly, albeit in rhetorical fashion, Wimsatt (2006: 461) remarks: "Memory – a property of molecules, neural circuits, tracts, hemispheres, brains-in-vats, embodied socialized enculturated beings, or institutions?" It is unmotivated to prohibit *a priori* the ascription of memory to both wholes and their parts, given that it is in fact being so ascribed by scientists across many fields or is clearly considered an empirical possibility that can help explain human and nonhuman behavior rather than provoke explanatory failure.

The homuncular functionalist might respond to such examples by saying that (e.g.) real memory is only being ascribed at the personal or whole level. But this response risks trivializing the view: whenever a cognitive capacity is apparently being ascribed at a subpersonal level, the homuncular functionalist can always dismiss it as not real. As a naturalistic theory, homuncular functionalism should not take for granted definitions of cognitive capacities, or interpretations of cognitive ascriptions, that entail that the theory cannot be falsified by evidence from the relevant sciences. Yet while the homuncular functionalist must posit such conceptual changes whenever the same word-forms are used at both personal and subpersonal levels, her reinterpretation of their meaning cannot be the default interpretation given that multi-level ascriptions of capacities in mechanistic explanations are often assumed to be conceptually continuous. For example, "rotates" picks out the same capacity exhibited by planets, ballerinas, and atomic nuclei even if each rotates in its own way.

But what then of Bennett and Hacker's (2003) mereological fallacy? This is the fallacy of using psychological terms to ascribe psychological capacities to parts when the very meaning of these terms makes such ascriptions nonsensical (according to Hacker's Wittgensteinian logico-grammatical orthodoxy). Bennett and Hacker explicitly argue that it is a *conceptual* mistake to ascribe (fully) cognitive capacities to parts. Dennett himself (Bennett et al. 2007: 88–89) responds to this fallacy on behalf of the homuncular functionalist: when the terms are used to ascribe capacities to parts, they ascribe "hemi-demi-semi-proto-quasi-pseudo" cognitive capacities (that is, homunculi) to the parts. As a result, no mereological fallacy is committed.

Dennett's response illustrates the point made above. He makes explicit the homuncular functionalist's need to posit conceptual change to explain away psychological ascriptions at subpersonal levels. Escape from the mereological fallacy comes at the cost of imposing *a priori* conceptual constraints of its own. (If it were a theory of meaning, one might say this *just is* the theory.) The problem remains that no such conceptual change can be taken for granted in the light of general scientific ascriptive and explanatory practice. Even if (*pace* Dennett) the psychological terms *are* being used to ascribe the same cognitive capacities to the parts, we have no reason to think that doing so runs afoul of any epistemic worries. Absent this epistemic motivation, the straightforward response to the mereological fallacy is to argue that psychological predicates are not conceptually barred from being ascribed sensically to the parts.[4]

This yields a puzzle. The ascription of the same capacities of wholes to their parts within a mechanistic explanation is not epistemically problematic in general. To the contrary, it is par for the course. Nor does scientific psychology appear to care about the homuncular functionalist constraints. Perhaps reprising *objects* at the level of parts really is epistemically idle: maybe one can't, on pain of epistemic idleness, build a dog using tiny dogs, rather than cells, as parts. But the epistemic idleness of this repetition for objects, even if true, clearly does not extend to capacities, not in general science and not in contemporary psychology.

An obvious solution to the puzzle is to show that psychological capacities are exceptions to the rule for adequate mechanistic explanations. Maybe there

[4] In Figdor (2018), I argue at length against their view and against the mereological fallacy. For present purposes, the homuncular functionalist's response is all that matters.

is something about *the mind* that makes repetition of *psychological* capacities at the level of parts invariably explanatorily pernicious.

This defense of homuncular functionalism might start from the idea that repeating a capacity of the whole at the level of parts would leave the explanation incomplete. But while incomplete explanations are relatively undesirable, they are not circular, idle, or question-begging merely because they are incomplete. The homuncular functionalist might add that in this case the mystery of the mind at the level of the whole would simply be repeated at the levels of the parts. Since nothing will have been done to dispel this mystery, the purported explanation would be viciously idle and circular, as feared. In incomplete non-psychological mechanistic explanations, in contrast, we know in principle how to fill in the gaps even when the same capacities are ascribed to parts. It's the fact that *the mind itself* is *especially* mysterious that yields a special problem.

The response implicitly concedes that if we understood better what psychological capacities are, it wouldn't actually matter if they were ascribed to parts. After all, the ban on repetition is intended to address a worry about a lack of explanatory gain. Non-psychological mechanistic explanations reveal that lack of explanatory gain is distinct from repetition. Repetition provides explanatory gain in that a repeated operation or capacity ascription still fills in details of mechanistic explanations, but the gain it provides will remain incomplete without illumination of the capacity. So the basic, and very real, issue is how to illuminate cognitive capacities. Homunculi are posited specifically to provide that illumination, of course, but their explanatory contribution depends on their gradually increasing stupidity (or gradually decreasing intelligence) as the decomposition proceeds, and so is not independent of the issue of repetition. One might wonder how much illumination is actually provided by the homuncular functionalist metaphors for this process. Setting that issue aside, if we can illuminate the mind naturalistically yet independently of the issue of repetition, we will have a naturalistic framework that does not require either homunculi or making psychology an outlier among the sciences.

Consider why at least some non-mental capacities are non-mysterious and can contribute to mechanistic explanations by being ascribed to parts as well as wholes. For example, why is rotating not mysterious at any level at which it is ascribed? One might provide two complementary mystery-dispelling reasons. Perhaps there are others, although these appear to suffice. First, we have some prior understanding of what rotating is – we've played with spinning tops, watched ballerinas, operated drills, and so on. So when entities that are too far away, too big, or too small to observe unassisted are said to rotate, the ascribed activity is not wholly mysterious, whatever it is ascribed to. Call this the familiarity condition.

Second, we have equations of angular momentum that help us distinguish instances of rotating from non-instances, even if the items doing the rotating are as distinct and as complex in their own ways as planets, ballerinas, and atomic nuclei. We individuate rotating as a type using a scientifically accepted method for distinguishing different kinds of motion, and we use that method to pick out tokens of rotating independently of our parochial perspective on which things rotate. The way each item achieves or exhibits its capacity to rotate may differ, but that's not problematic. The equations help guide our ascriptions by providing a standard for determining when things we think are rotating really are and when they aren't. Call this the objective constraints condition.

In sum, armed with a scientifically accepted kind of evidence for and constraints on the ascription of rotating to an entity, plus some prior understanding of that capacity, there is no fear of perpetual mystery that might motivate a ban on ascribing rotating to parts within mechanistic explanations of the rotating of wholes.

In the case of psychological capacities, the familiarity condition is clearly met. We do have some familiarity with what these capacities are from our own case. There's nothing wrong with that; we need to start explaining the mind from somewhere. We don't also need to think we're infallible or that the mind is transparent to introspection.

What is new and significant is that the objective constraints condition is also being met. The contours of this type of illumination in psychology are now discernable through the use of mathematical models of cognitive capacities. The mathematical models I have in mind are those expressed using equations that formally describe empirical relationships the way mathematical models of non-psychological phenomena (such as the Hodgkin-Huxley model of the action potential) do. A psychological example of such a model is the drift-diffusion model (DDM) of two-choice decision-making, first proposed by Roger Ratcliff (1978) and subsequently elaborated, tested, and extended by Ratcliff and colleagues and other psychologists. (The TD model mentioned above is another.) This model proposes cognitive processes of evidence accumulation and assessment to a threshold, at which point a decision is made and a behavioral response is given. For example, subjects (often undergraduates) may be asked to press a key to report whether an image is of a house or a face, and the experimenter manipulates the clarity of the image. The model posits cognitive processes that mediate between the stimulus-response relationship in a way that captures the speed-accuracy tradeoff: the relation between the clarity of the evidence and the speed and accuracy of the subsequent responses. Given the same degree of noisiness in the stimuli, subjects make more mistakes (relative to benchmarks) when they are instructed to respond quickly and respond more slowly (ditto) when instructed to emphasize accuracy.

What is formalized by such models are the observable behavioral patterns of people (and some nonhuman species) from which we infer to a cognitive capacity or combination of them that can yield these patterns. This is crucial when we are interested in the capacities of entities that are not already in the cognitive club by hypothesis or general consent. For example, the DDM equations provide a standard scientific means to identify tokens of the posited decision-making processes independently of our parochial understanding of mind. In the case of undergraduates, the appropriateness of ascribing internal cognitive processes of accumulating and assessing evidence to a decision threshold is assumed. Human behavioral data was used to develop the model in the first place. But the DDM has also been used to examine whether fruit flies' decision-making is affected by a genetic flaw that also affects humans (Dasgupta et al. 2014). It was not given that the model could be used for fruit flies. It was a matter of empirical test. This fact grounds its objectivity: it is not up to us to decide where it may fit. We may not otherwise have evidence of a behavioral regularity or of its similarity to the human behavior on which we base cognitive ascriptions. Ordinary observation can even impede this recognition. For this reason, the model can be used to determine when something is behaving relevantly similarly to the way we do when we have made a simple decision and then act on it, even if the entity is not of a kind that we intuitively think of as being able to make decisions.

Of course it does not *follow* that the natural language expressions used to interpret the equations are being used in the same way when the models are extended successfully in new domains. But extension by means of formal models is part and parcel of the same scientific methodology by which we ascribe capacities such as rotating or oscillating across vastly different domains as well. The practice increasingly includes cognitive science and social science (e.g., Irvine 2016, Froese et al. 2014). Moreover, in contemporary network science, the use of the same models at multiple spatiotemporal scales is integrated into the practice of explaining the behavior of wholes in terms of the behavior of their parts (e.g. Alon 2007, Baronchelli et al. 2013). This is illumination of the sort Newtonian mechanics achieved between the terrestrial and celestial domains via one set of laws of motion that applied to both. Not incidentally, Newton's laws naturalized the celestial realm, whereas before it was mysterious and divine.

The fact that we know how to satisfy the objective constraints condition for psychological capacity ascriptions does not imply that we will not find other types of objective evidence. Cognitive neuroscientists are actively seeking neural activation patterns that will enable us to "mind-read" by using neural behavior as the basis for our inferences (Poldrack 2006; Roskies 2014). Such evidence could complement behavioral evidence, and in cases of conflict the outcome for ascriptions will depend on other factors. It also does not imply that a single model for each intuitively individuated cognitive capacity will suffice. Naturalizing the mind will require confronting at least three distinct sources of complexity in traditional psychological ascriptions: distinguishing and relating various types and subtypes of cognitive processes; distinguishing the non-epistemic and epistemic goals of our traditional cognitive-ascriptive practices; and articulating differences in the grounds and types of meaning adjustment as word-forms are used in new domains. These factors will have to be disentangled as naturalization proceeds. The present point is that mathematical models show how we can begin to satisfy the objective constraints condition for the mind. If we have evidential means to ground cognitive ascriptions to parts, and the ascriptions are being made on these grounds, and the evidential means and ascriptive practices follow standard scientific canons, and all of this violates the homuncular functionalist constraints on an adequate naturalistic explanation of mind, *tant pis* for homuncular functionalism.

Dennett describes discharging as the point at which there are no questions about intelligence being begged. Cognitive capacities are no doubt more complex than rotating. But what would be missing if at every level at which a cognitive capacity were ascribed on the basis of a model, we could then explain mechanistically how the entities at that level realized that capacity? The explanation would be incomplete until we filled in these details, but there is no reason to think we have perpetuated mental mystery. Naturalization without homunculi implies that psychological capacities may not be eliminable. But explanation does not require elimination. It requires eliminating mystery. Only if we think the psychological is essentially mysterious does it follow that eliminating the mystery of the psychological requires eliminating the psychological. Oddly enough, the homuncular functionalist method implies that psychological concepts and the capacities they pick out are not just mysterious now, but essentially so.

Note that the phrase "psychological model" (or "cognitive model") has a longer history and wider scope that includes more than just the mathematical models discussed above. The phrase also encompasses boxologies or other informal functional analyses of cognitive processing, as well as computational models of cognition. The latter are connectionist (or hybrid symbolic-connectionist) networks used as stand-ins for neural networks and their activation patterns; they are models of possible realizations of cognitive capacities by neuron-possessing creatures that by hypothesis already belong to the cognitive club. Note that these networks may also be called mathematical models – I am not seizing the label but merely using it in this paper to be precise about the sort of psychological model that can play the evidential role for psychological ascriptions that (e.g.) equations of angular momentum do for ascriptions of specific motions.

The important difference between these types of models and the mathematical models of interest here is that one cannot use these other types of models on their own to guide and constrain psychological ascriptions across domains. They may satisfy other explanatory purposes, but not the purpose of determining which things have cognitive capacities. Boxologies do not include operationalized behavioral signatures of the posited capacities, although they can be augmented with them. If formal, this would supplement the boxologies with the sort of evidence that mathematical models contribute; if the behavioral patterns are not formalized, the augmentation would fall short of providing objective constraints on psychological ascriptions to non-human domains.

Artificial neural networks do link input with output, but the operationalization consists of assigning numerical values (vectors) to whatever the inputs and outputs happen to be. The evidential work of interest here is done by identifying the stimuli-response pairs that the input and output vectors represent. Again, if these relationships are not formalized, we would not have an objective means to extend the network to new domains, even if we set aside the fact that cognitive capacities in entities without brains (such as plants or slime molds) are beyond the scope of these models from the start. In addition, connectionist networks are used widely outside psychology. Whether what they represent is a psychological process at all is not determined by anything intrinsic to the network. If or when future artificial network designs make such identification possible, they would be an additional tool for investigating internal evidence of cognition, alongside the "mind-reading" research mentioned above.

In sum, we now have the tools to overcome the evidential gap for ascribing psychological ascriptions to nonhumans in an objective manner. Mathematical models of cognitive capacities rely on our normal understanding of the capacities being modeled and so satisfy the familiarity condition for illumination. When patterns of behavior of humans is captured formally, we can see if behavioral data from nonhumans satisfies the model, whatever we may think of those nonhumans or their behavior. Formalization provides clear constraints on when we are entitled to ascribe to an entity the capacities that we ascribe to humans using psychological language. This procedure satisfies the objective constraints condition of illumination. Ascriptions of harmonic oscillation or rotation to entities at multiple scales follow the same procedure. In the case of the mind, unlike those cases, the ascriptions are made by inference from the observed patterns. But that is true of our ascriptions of cognitive capacities to each other. Finally, these possible extensions include entities that are parts of wholes to which we also ascribe the capacities on the basis of the behavior captured by the same model.

Armed with this alternative, we simply do not need homunculi to naturalize the mind. The dormitive virtues lost their explanatory power because virtues no longer had explanatory force within the mechanistic explanatory framework. Modern-day homunculi can go the way of the dormitive virtues. In contemporary scientific psychology, they too are now idle, along with the discharging requirement associated with them.

4 Model-based Naturalization, Mechanisms, and Autonomy

In the last section I distinguished among types of psychological models in a way that reflects the present paper's concern about the viability of homuncular functionalism give contemporary explanatory practices in the cognitive sciences. In this section I will consider how the model-based framework for naturalization intersects with two current debates regarding scientific explanation in general and psychological explanation in particular. Their common starting point is mechanistic explanation. First, do laws or models provide explanations at all? Second, is psychological explanation autonomous from neuroscientific explanation? I will show how the view is neutral regarding these debates. The questions they raise remain outstanding even if we reject homuncular functionalism in favor of model-based naturalization.

First, do models or laws *explain*? This is a long-running debate between the Hempelian covering-law model of explanation and mechanistic explanation as the two basic conceptions of scientific explanation. For example, Craver 2006 calls the Hodgkin-Huxley equations "phenomenal models" that can be used for prediction, like covering laws, but they do not *explain* because they do not detail the mechanisms. Presumably the DDM and other mathematical models of cognition also count as phenomenal models and so would also not explain for the same reason. More recently, mechanists have proposed that psychological models are mechanism sketches, which are incomplete mechanistic explanations or when filled in with the details turn into mechanistic explanations (Piccinini and Craver 2012). This position accepts that models play a role in scientific explanations, but ties their ability to explain to that of mechanisms. Opponents claim that laws or phenomenal models provide explanations independently of whatever explanatory work is provided by detailing the mechanisms (e.g. Batterman and Rice 2014, Chirimuuta 2014).

Relative to this debate, it is sufficient for my purposes to claim merely that laws and models contribute to explanation, for the feature of interest here is their applicability to multiple levels of mechanisms. Phenomenally adequate cognitive models combine evidence of patterns of behavior with a familiar psychological conceptual framework that provides some understanding of the capacities that may be ascribed at multiple levels. It is a further question whether the models' contribution at any level to which they apply is due to their own explanatory power or because they function as mechanism sketches. They certainly do provide constraints on capacity ascriptions, non-cognitive or cognitive. Having such objective individuation criteria is an important advance in understanding, whether we want to consider this advance an explanation or not. Models and mechanisms may simply have a symbiotic relationship. As Sober (1982: 421–422) remarks with regard to laws, if we are told an organism digests in part because parasites in it digest, "we now want to know what laws govern the way organisms obtain energy from their environments, and how those laws apply simultaneously to hosts and the parasites they house."

Homuncular functionalism, in contrast, occupies an awkward position in relation to this debate. It holds that an activity of a part does no explanatory work if it is the same type of activity as the activity of the whole to which the part belongs and which the part-level activity is supposed to help explain. But it also holds that an activity of a part *can* do explanatory work if it is related to the capacity of the whole as being lesser on a continuum of similarity that gets whittled down to nothing. In other words, what makes a homunculus dissimilar makes it explanatory and what makes it similar perpetuates explanatory idleness. Neither law-based nor mechanistic explanation affirms these claims. For example, the equations of angular momentum unify a planet's rotating and an atomic nucleus' rotating. The laws reveal an important similarity across domains, and for some philosophers (and scientists) this unity is the essence of explanation. For mechanists, it simply doesn't matter if planets and their atomic nuclei both rotate, for the explanation is provided by showing *how* they rotate, not that the mechanisms are relevantly similar or different. An adequate answer to the 'how' question neither requires nor rules out multiple realizability of the explanandum phenomenon.

Second, are psychological explanations in some important sense autonomous from neuroscientific ones? On the assumption that the latter are mechanistic, this question becomes a special case of the first debate, although the question of the autonomy of psychology from neuroscience predates the rise of the new mechanism. In contemporary terms, the claim that cognitive models are mechanism sketches (Piccinini and Craver op.cit.) or that they need to be mapped to mechanisms to have explanatory power (Kaplan and Craver 2011) is a way to argue that psychological explanations are not autonomous. In contrast, Weiskopf (2011: 322–23) distinguishes between psychological models that aim to explain psychological phenomena in semantic, intentional, or representational terms, and those that aim to explain psychological phenomena in non-psychological (e.g., neurobiological) terms. The first group includes connectionist networks whose states, while subpersonal, are interpreted in representational terms. Focusing on the first group, Weiskopf argues that models in this group are not mechanistic because their components don't correspond to the parts of the modeled system in any straightforward way. This would be a way to defend the autonomy of psychological explanation.

The model-based view of naturalization is neutral regarding the autonomy debate. It allows for psychological ascriptions at personal and subpersonal levels. Whether these models are mechanistic, explanatory on their own, or autonomous from neuroscientific models or mechanisms are further issues – the naturalization project does not depend on how they are resolved. Homuncular functionalism, in contrast, implies that ultimately psychology can't be autonomous. The view straddles Weiskopf's categories: the explanation starts out in the first category but ends up in the second. So even if psychology starts out autonomous, the upshot of the explanatory framework is to undermine that autonomy step by step even if that is not its goal.

As a final issue, model-based naturalization suggests that the traditional debate between reductive vs. non-reductive physicalism must be reconfigured for a scientific context in which the same formal structures are employed at multiple scales. An implicit assumption of the traditional debate is that these are distinct conceptual schemes whose distinctness is in part tied to particular levels. The gradualist framework of homunculi takes on this implicit assumption. The model-based view of naturalization makes reduction a levels-relative (and not just capacity-relative) affair. If psychology reduces to neuroscience at one level, this result cannot automatically be generalized to all levels. Moreover, if assigning entities to levels is difficult, assigning capacities to levels will also be difficult (Stinson 2016). Even the most favorable cases of reduction may require significant simplification of the phenomena.

5 Concluding Remarks: Naturalization Without Homunculi

The contemporary explanatory context for psychology is one in which models of capacities are apt in principle for use at any level in a mechanistic hierarchy. If the models apply to wholes and their parts, so be it. Positing ever-stupider homunculi in this explanatory context is like insisting that only ballerinas really rotate even though the same equations of angular momentum apply to their atoms. We don't need stupider capacities at each level. To the contrary, therein lies the perpetuation of mystery.

It is difficult to identify what homuncular functionalism gets right, if anything. Perhaps it is just the idea (surely not unique to homuncular functionalism) that there is a special epistemic problem involved in naturalizing the mind. But homuncular functionalism did not identify the problem correctly. The unique challenge faced in psychology is the problem of distilling from our homegrown, first-personal, often introspective understanding of the psychological those features that are specific to humans but contingent to possession of the capacity. Until recently, we had no idea how to gain a non-anthropocentric perspective on the mind. Now we do. Models can help us distinguish those aspects of a cognitive capacity that are matters of human realization and those which are arguably what it is to have that capacity. Where subpersonal entities differ from humans (or other wholes with minds) need not make any difference to their possession of the capacity, just as the fact that planets and atomic nuclei lack arms and legs does not deprive them of the capacity to rotate.

Have I argued for a different type of homunculus? No – the proposal is not that neurons are little men. The proposal is that parts of men can, in principle, do what men do, and that we have the empirical tools to discover whether they do or not. Have I argued that consciousness can be explained with non-intrinsic properties? No – I don't know how consciousness will end up being explained. Have I shown that the mind can be explained naturalistically? No – I've taken for granted here that it can be. What I have shown is this: if the mind can be explained naturalistically, we have an alternative to the homuncular functionalist framework that is a natural fit with the way many accepted explanations in the sciences are actually formulated, and with the way contemporary cognitive researchers are pursuing their explanatory goals.

Acknowledgments:

I wish to thank audiences at the 8th Quadrennial Fellows Conference (Lund, Sweden) and Exploring the Undermind conference (University of Edinburgh) in July 2016, and the British Society for Philosophy of Science Annual Meeting and Aristotelean Society/Mind Joint Sessions (Edinburgh) in July 2017 for comments on earlier versions of this paper, and two anonymous referees for this journal of the revised paper.

References

Alon, U. (2007). Network Motifs: theory and experimental approaches. *Nature Reviews Genetics* 8: 450–461.

Baronchelli, A., R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, and M. Christiansen (2013). Networks in cognitive science. *Trends in Cognitive Sciences* 17 (7): 348–360.

Bechtel, W. (2008). *Mental Mechanisms*. New York and Oxon: L. Erlbaum.

Bechtel, W. and A. Abrahamsen (2005). Explanation: a mechanist alternative. *Studies in the History and Philosophy of Biology and Biomedical Sciences* 36: 426–441.

Bennett, M. and P. Hacker (2003). *Philosophical Foundations of Neuroscience*. Malden, MA and Oxford: Blackwell.

Bennett, M., D. Dennett, P. Hacker, and J. Searle (2007). *Neuroscience & Philosophy: Brain, Mind, and Language.* New York: Columbia University Press.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Clarendon.

Chirimuuta, M. (2014). Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese* 191: 127–153.

Craver, C. (2006). When Mechanistic Models Explain. *Synthese* 153: 355–76.

Craver, C. (2007). *Explaining the Brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: OUP.

Craver, C. and J. Tabery (2017). Mechanisms in Science. *The Stanford Encyclopedia of Philosophy* (Spring 2017 edition), E. Zalta, ed., URL = https://plato.stanford.edu/entries/science-mechanisms/

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge: MIT.

Dasgupta, S., C. Howcroft Ferreira, G. Meisenböck (2014). FoxP influences the speed and accuracy of a perceptual decision in Drosophila. *Science* 344 (6186): 901-04.

Dennett, D. (1975). Why the Law of Effect Will Not Go Away. *Journal of the Theory of Social Behavior* 5: 169–187.

Figdor, C. (2018). *Pieces of Mind: The proper domain of psychological predicates.* Oxford and New York: Oxford University Press.

Fodor, J. (1968). *Psychological Explanation*. New York: Random House.

Froese, T., C. Gershenson, and L. Manzanilla (2014). Can Government Be Self-Organized? A mathematical model of the collective social organization of ancient Teotihuacan, Central Mexico. *PLoS One* 9 (10): e109966.

Irvine, E. (2016). Model-Based Theorizing in Cognitive Neuroscience. *British Journal for the Philosophy of Science* 67: 143–168.

Kaplan, D. and C. Craver (2011). The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philosophy of Science* 78 (4): 601–627.

La Mettrie, J. (1748). *Man—Machine*. Trans. Jonathan Bennett, 2017. Downloaded from http://www.earlymoderntexts.com/assets/pdfs/lamettrie1748.pdf.

Leibniz, G. (1714). *The Principles of Philosophy, or the Monadology.* Sec. 17. Trans. Ariew and Garber (1989) *Leibniz: Philosophical Essays* (Indianapolis and Cambridge: Hackett): 215.

Lycan, W. (1991). Homuncular Functionalism Meet PDP. In Ramsey, Stich and Rumelhart, eds., *Philosophy and Connectionist Theory.* L. Erlbaum: 259–86.

Maienschein, J. (2017). Epigenesis and Preformationism. *The Stanford Encyclopedia of Philosophy* (Spring 2017 edition), E. Zalta, ed., URL = https://plato.stanford.edu/entries/epigenesis/

Machamer, P., L. Darden, and C. Craver (2000). Thinking About Mechanisms. *Philosophy of Science* 67 (1): 1–25.

Margolis, J. (1980). The Trouble With Homuncular Theories. *Philosophy of Science* 47 (2): 244–59.

Piccinini, G. and C. Craver (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183: 283–311.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* 10 (2): 59–63.

Roskies, A. (2014). Mindreading and privacy. In M. Gazzaniga, ed., *The New Cognitive Neurosciences* (Cambridge, MA: MIT Press): 1003–11.

Sober, E. (1982). Why must homunculi be so stupid? *Mind* 91: 420–422.

Stinson, C. (2016). Mechanisms in psychology: ripping nature at its seams. *Synthese* 193: 1585–1614.

Suri, R. and W. Schultz (2001). Temporal Difference Model Reproduces Anticipatory Neural Activity. *Neural Computation* 13: 841–82.

Sutton, R. and A. Barto (1981). Toward a Modern Theory of Adaptive Networks: Expectation and Prediction. *Psychological Review* 88 (2): 135–170.

Weiskopf, D. (2011). Models and mechanisms in psychological explanation. *Synthese* 183 (3): 313–338.

Wimsatt, W. (2006). Reductionism and Its Heuristics: making methodological reductionism honest. *Synthese* 151: 445–75.

Jelena Issajeva
Tallinn University of Technology,
Estonia
Akadeemia tee 3, 12611 Tallinn, Estonia
E-mail: jelena.issajeva@gmail.com
Ahti-Veikko Pietarinen
Tallinn University of Technology, Estonia
Akadeemia tee 3, 12611 Tallinn, Estonia

Original Scientific Paper
UDK 165.24:159.954/.955

THE HETEROGENOUS AND DYNAMIC NATURE OF MENTAL IMAGES: An empirical study

Abstract. *This article addresses the problem of the nature of mental imagery from a new perspective. It suggests that sign-theoretical approach as elaborated by C. S. Peirce can give a better and more comprehensive explanation of mental imagery. Our empirical findings follow the methodology of cognitive semiotics and they show that (i) properties of mental images are heterogenous in nature; (ii) properties of mental images are dependent on the characteristics of object-stimulus; (iii) properties of mental images are dependent on individual differences in imaginary capacities. This suggests that, contrary to representational accounts, mental imagery is not based on one dominant representational format. Imagery constitutes a complex system of signs consisting of several sign elements and dynamic relations. A sign-theoretical account may give a better explanation of the nature of mental imagery, as it accommodates heterogenous evidence from this experiment.*

Keywords: *mental imagery, representation, experimental semiotics, theory of signs, Peirce.*

1. Introduction

Since the ‘cognitive revolution’ the question about the nature of mental imagery (MI) remains one of the most debated ones in cognitive sciences and philosophy of mind. This article presents new empirical evidence on the matter that follows the methodology of cognitive semiotics. The experimental results showed that (i) properties of mental images are heterogenous in nature; (ii) properties of mental images are dependent on characteristics of object-stimulus; (ii) properties of mental images are dependent on individual differences in imaginary capacities. These results conform with empirical data in neuroscience (Bartolomeo 2008; Moro et al. 2008; Dulin et al. 2008), which claim that one dominant representational account does not adequately

doi: 10.5937/BPA1831057I

explain MI. Supported by empirical findings, this paper argues that mental imagery is better explained in terms of sign theory as proposed by C. S. Peirce.

What is MI? Traditionally, two dominant rival theories have been proposed, (quasi-)pictorial and propositional. According to the (quasi-) pictorial theory, mental images are picture-like representations in the mind (Kosslyn 1978, 1980, 1994; Pinker and Finke 1980; Finke, Pinker and Farah 1989). Proponents of propositional theory, on the other hand, claim that MI constitutes verbal representations or language-like descriptions (Pylyshyn 1973, 1981, 2002, 2003; Fodor 1975, 1990). The controversy between the two constitutes the Mental Imagery Debate.[1] It is common to understand both rivals in the framework of computational-representational paradigm[2] of mind, which implies that all mental states are products of mental computation. In this context, representational theory is focused on a search of one dominant format or code[3] that underlies mental imagery, as well as other mental states.

Adherents of both representational theories have deployed empirical methods to prove their respective claims. Kosslyn and colleagues have experimentally shown that mental imagery (MI) has certain spatial and picture-like properties (size, colors, shapes, dimensions, distances, etc.) and thus concluded that images are most likely pictorial representations (Kosslyn 1980, 1988, 1994; Kosslyn et al. 2006; Shepard and Metzler 1971; Pinker and Finke 1980; Shepard and Cooper 1982; Slotnick et al. 2005). In contrast, Pylyshyn and colleagues (Fodor 1975, 1990; Slezak 1990, 1991, 1995) argued that there is substantial empirical evidence to think of images as being descriptions formulated in language(-like) terms rather than pictures (see for example Fodor's "Language of Thought" hypothesis, 1975).

Despite persistent ambiguities of data on the matter, Imagery Debate was claimed to be solved in favor of (quasi-)pictorial theory of MI (Kosslyn 1994; Pearson and Kosslyn 2015). However, in the light of new experimental methods and results the previous long-standing theories of MI have been reconsidered. Research shows that there are significant difficulties in the representationalist (either pictorial or propositional) approach, some of which are inherent to the representational-computational paradigm (Milikan 1984; von Eckardt 1993;

[1] On the Imagery Debate and details on the theories of MI, see Thomas 2010, 2014; Kosslyn 1980, 1988, 1994; Kosslyn et al. 2006; Pearson and Kosslyn 2015; Pylyshyn 1973, 1981, 2002, 2003; Tye 1991.

[2] On representational-computational paradigm of mind see Van Gelder 1995; Clapin 2002; Marr 2006.

[3] The discussion of whether there are two mental codes that underlie our mental states is called dual-common coding debate. The latter stems from Alan Paivio and his work on memory and learning effects (Paivio 1971, 1986). Noteworthy, dual/common coding debate is different from Imagery Debate (which is also called analog/propositional debate), but has often been confused. Dual-common coding debate focuses on whether we learn and memorize information by using one mental code or another. The analog/propositional debate, in contrast, investigates the nature of MI. It is the latter, which is the focus of present study.

Bechtel 1998; Knuutila 2005, 2011). First, under similar experimental settings mental imagery can exhibit (at least) both types of properties – verbal and pictorial (Anderson 1978; Pylyshyn 2002; Ganis 2013). Also, most empirical results on MI yield multiple interpretations (for a detailed discussion of explanations of the experimental results, see Pylyshyn 2002). Most replicated experiments on MI often show differing results (Pylyshyn 1981, 2002; Slezak 1990, 1991; Chambers and Reisberg 1985; Rock, Wheeler and Tudor 1989). There is also a significant amount of empirical evidence proving the existence of motor, tactile, auditory properties of mental imagery (for details, see Lacey and Lawson 2013; Keller 2012; Pascual-Leone et al. 1995; Plessinger 2007; Richardson 1995; Gregg and Clark 2007; Schimdt et al. 2014). All this yields to the conclusion that current empirical data cannot be fully accommodated either by (quasi-)pictorial or by propositional accounts of MI. Maybe understanding of mental imagery cannot be restricted to the dichotomy verbal-pictorial and the search of one dominant format[4] of MI is misleading. What is mental imagery really like? Is it a picture in the brain, some propositional or verbal string of language-like characteristics, or something else still?

Most of the novel approaches have emerged in this context, such as enactive and attention-based quantification theories.[5] Both attempt answering the question of what the true nature of mental imagery is (Thomas 2010; Sima 2011). According to the enactivist approach, mental imagery is a mental capacity of an active cognitive search of information in the absence of the actual perceptual stimulus (Thomas 2009: 454–455). Although enactivism is a representational account, it encounters problems such as vagueness of explanation of the nature of MI and inability to explain deep complexity and multiplicity of properties of images. Yet another alternative account – attention-based quantification theory – explains imagery in terms of attentional processes that quantify spatial and visual information by operating upon two working memory structures, namely Qualitative Spatial Representation (QSR) and Visuo-spatial Attention Window (VSAW) (Sima 2011: 2880). The attention-based quantification theory tries to integrate memory and attention to explain MI, but it relies on qualitative representations and encounters the same difficulties as other representational theories. In sum, increasing diversity of alternative theories has not solved the question.

[4] By the term "representational format" we mean internal structure of the mental image, or its "cognitive architecture". We use the term "cognitive architecture" largely in the sense of Z. Pylyshyn (2002), namely to mean the underlying structure of MI, that is "properties and mechanisms [that] are *intrinsic* to, or *constitutive* of having and using mental images" (Pylyshyn 2002: 159, original emphasis). Noteworthy that we do not take "representational format" to mean phenomenal modality.

[5] Alternative theories – enactivism and attention based quantification theory – are relatively unpopular views in solving the issues of MI. Dominant accounts remain representational. Thus, current empirical study was designed to test the consistency of most dominant representational accounts on MI.

This article suggests a novel semiotic approach to address the question. In particular, it argues that Peirce's theory of signs as proposed applies well to the analysis of mental imagery and that it can give a coherent explanation of diverse empirical data. We begin with a brief analysis of the theoretical premises of the theory of signs as contrasted with traditional representational accounts. Section 3 describes the experimental design, hypotheses, experimental methodology and procedure. Section 4 provides the results of the experiment. Section 5 is the discussion of the results and their analysis.

2. Theory of signs and MI

Peirce's theory of signs (or semeiotic) is an account of signification, reference and meaning (Pietarinen 2015). There are several formulations of theories of signs (see for example Saussure 1983; Morris 1938, 1946, 1964), but Peirce's account is distinctive for its "breadth and complexity" (Atkin 2017: 1). It interprets MI[6] as a complex *system of signs,* which consists of three elements – representamen, object and interpretant – and is characterized by dynamic and flexible relations between these elements (Issayeva 2015; Pietarinen 2012). Such an approach begins with the premise that the mind is of a signifying nature. In particular, the human mind is a sign-producing and sign-interpreting system, characterized by the semiotic processes of signification, i.e. by dynamic, changing and context-dependent processes that create signs and manipulate them. Peirce associates cognition with signs. According to Peirce, all our mental states are signs: "we think only in signs" (Peirce 1994, CP 2.302) "a theory of experience, a theory of consciousness" (Zeman 2014: 1). The human mind constitutes "a historically existing continuum of interpretants (which are signs)", i.e. the the history of signification of objects in one's mind (Zeman 2014: 2). Peirce's famous claim was that "man is a sign" (Peirce 1994, CP 5.314; Peirce 1998, EP 1:54), a person consists of her own thinking, and since all thoughts are in signs, a person is a historical series of signs.

But what does it mean to say that MI is a sign system? First, mental imagery has a signifying nature. MI shares the same structure and features with a sign in the human mind. Just as a sign is defined as "something which stands to somebody for something in some respect or capacity" (Peirce 1994, CP 2.228) in the same way a MI can be legitimately characterized. Thus, just as a sign, MI is comprised of three main elements: representamen, object and interpretant (Peirce 1994, CP 2.228; Peirce 1998, 478). MI also has a

[6] Noteworthy that sign-theoretical definition of MI is twofold. According to Peirce imagery can be understood 1) in a narrow sense meaning the representamen element or something that stands for something, or 2) in a broader sense meaning a mental entity consisting of three elements and constituting a signifying whole, i.e. a system of signs. We take mental imagery as a faculty to mean the second. Though, both definitions can be met in the discussion.

representamen, which is an element that stands for some object or event. MI has necessarily an object, which it signifies. And it has an interpretant or the meaning that holds between representamen and its object. Hence, every mental image signifies something.

Second, to say that mental imagery is a sign system means that MI is guided by dynamic, context-dependent signifying relations between its elements. According to the sign-theoretical account, relations between elements of the sign are dynamic, i.e. they continuously develop and change their characteristics dependent on various factors. As Floyd Merrell (2001) puts it: "signs simply cannot stand still" (Merrell 2001: 37). Similarly, mental images are not stable or fixed, but are rather of dynamic nature. MI evolve and continuously develop under the influence of both internal (e.g. subjective memory, experience and dispositions) and external (e.g. changes in language, objects' features and new knowledge) factors. The latter entails that mental images are dependent on the context, where they were produced, as well as on the subject, who produced or interpreted an image. The changes in the environment significantly influence both the relations and characteristics of the MI elements. Shortly, context as well as personal experience and cognitive dispositions matter. These influence the whole process of signification. Sign-theoretical account as proposed by Peirce can accommodate these features. The dynamics, openness and flexibility of the triadic relations allows to explanation of divergent and changing properties of images under the umbrella of one sign-theoretical account.

To sum up, a sign-theoretical approach towards the explanation of mental imagery yields that a mental imagery is a sign system, which consists of three main relata – a signifying-vehicle or representamen, an object and an interpretant – and is characterized by dynamic, context-dependent semiotic relations between them. Together they – a set of sign-elements and semiotic relations – constitute an interconnected network that works together as a whole. Schematically, these assumptions can be depicted in the following way:

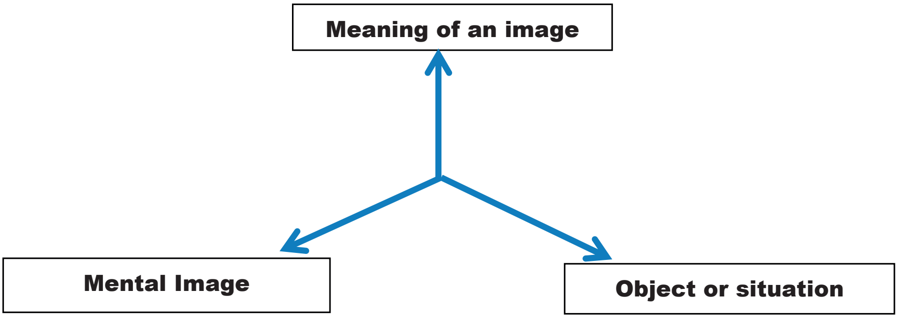


Figure I: Structure of a mental image

Based on the triadic structure of a sign, Peirce elaborated a fine-grained classification of signs. Each of the three sign elements – representamen, object and interpretant – can be further analysed and divided into three subtypes (Peirce 1994, CP 2.243). In relation to representamen a sign can be divided into Qualisigns, Sinsigns and Legisigns, in relation to its object to Icons, Indexes and Symbols, and finally the third division, Rhemes, Dicisigns and Arguments, is related to the analysis of interpretant (Peirce 1994, CP 2.244–2.252). Together these three trichotomies give rise to ten classes of signs. Applied to MI, classification guides a comprehensive investigation of a mental image and each of its relata, and makes Peirce's semiotics a promising method for an in-depth analysis of MI.

A sign-theoretical approach, thus, might provide a new perspective to MI. It can potentially overcome traditional controversies and limits of the representational accounts and might eventually give a sound and coherent explanation of the MI and its relata.

3. Methods

How to test these assumptions? How can we prove that mental imagery indeed can be legitimately interpreted as a sign system?

To test this, we have conducted a test using the methodology of cognitive semiotics and experimental philosophy. The choice of the method of investigation was not arbitrary. First, semiotics is the science that studies signs and their use. In particular, methods of experimental cognitive semiotics and examples of cognitive task design offer a unique way to test the production of a sign in its dynamics and track the most fundamental features of the signification process.

Second, MI constitutes a complex theoretical and interdisciplinary problem with a long tradition in philosophy of mind. Thus the research methods from philosophy sharpen the theoretical hypotheses and our experimental design. In particular, the methodology of experimental philosophy was employed to empirically investigate MI. Experimental philosophy approaches philosophical problems from empirical perspectives. This is of a special value in approaching the debate about MI.

Finally, experimental methods[7] conform to the investigation of imagery in cognitive psychology. The latter has resulted in divergent empirical data

[7] Although current study uses mostly qualitative research methods and does not rely on the brain scanning techniques (fMRI and EEG), the results of our study are still taken to be contributive to the discussion of MI. There are a series of high-powered qualitative research that gives important results on the functions of images without brain scanning and physiological response potential techniques. Actually, some of the classic and most influential experiments on MI (e.g. Perky 1910; Shepard and Metzler 1971; Paivio 1971) are qualitative in nature. Besides, traditional representationalist experimentations on MI (mental scanning, mental mapping and mental paper folding), while based on EEG or

regarding the nature of MI. Hence, these experimental methods seem appropriate and they take into consideration the relevant previous research. Thus, to show the applicability of the sign-theoretical approach towards the investigation of MI, the relevance of its results to the understanding of the nature and function of MI, as well as its correspondence to the previous research, the experimental method was chosen as the most suitable way to test whether mental imagery shares the same characteristics with a sign.

The experimental design is based on the standard methods and cognitive tasks used in cognitive semiotics and in experimental philosophy. The experimentation began with the short introductory pre-test questionnaire to check the statistically relevant information about age, nationality, cultural and educational backgrounds of the participants. The pre-test was followed by an actual experiment that consisted of three different cognitive tasks. An experiment was finished by the Psi-Q after-test (The Plymouth Sensory Imagery Questionnaire).

The latter constitutes a well-known test on evaluation of imaginary capacities – its vividness and intensity – that was elaborated by psychologists at Plymouth University (Andrade et al. 2013). The essential advantage of the Psi-Q test as compared to other similar questionnaires[8] is its sensitivity to images across a wide range of modalities: vision, sound, smell, taste, touch, bodily sensation and emotional feeling. This allows to test individual differences in imaginary capacities in more detail. For this reason the Psi-Q test was chosen to measure individual imagery capacities. Finally, the data gathered was analysed using the methods of descriptive statistics –SPSS and R-studio digital services.

3.1. Experimental hypotheses

Based on the theoretical premises of a theory of signs the following experimental hypotheses were formulated. The **main theoretical hypothesis** is that mental imagery can be legitimately viewed as a system of signs:

1) MI shares the same structure with a sign. In other words, MI has an object, interpretant and representamen.
2) MI is formed in a semiotic process, i.e. inside a network of the signifying relations. The relations between MI's elements define the particular properties of the final image produced.

fMRI, also rely on qualitative methods, including introspection and self-reports. Thus, we believe that our choice of experimental methodology is justified and conforms both to the standards and the practice of the methodological choices used to investigate MI.

[8] There are several tests to evaluate vividness of MI: Betts Questionnaire upon Mental Imagery (QMI; Betts 1909; Sheehan 1967), Marks' Vividness of Visual Imagery Questionnaire (VVIQ and VVIQ2; Marks 1973), Gordon's Test of Visual Imagery Control (TVIC; Gordon 1949). The Psi-Q test was chosen before other alternatives, because it allows to evaluate not only the visual MI, but vivdness of images across all sense-modalities.

To test these the triadic structure of an image was manipulated to uncover the potential correlation between the properties of imagined object and the properties of the final image produced. As a result theoretical sub-hypotheses 1 and 2 were simplified into the following experimental hypotheses:

> **H0:** Mental image has the same or similar characteristics, regardless of the characteristics of an object.
>
> **Ha**: Mental image has different characteristics. Particular properties of an object influence the characteristics of an image formed to present this object.

In order to analyze these hypotheses, cognitive experimentation was divided into three tasks given to each participant: pictorial, verbal and diagrammatic. Following Peirce's typology of signs, such a task division was chosen to represent distinctive differences in object-stimulus that were supposed to influence a final image. The judgement about statistical significance of the test results will be made on the basis of significance level, the value of which for the sake of the current experiment is taken to be 0.05 (i.e. $\alpha = 0.05$). The choice of the significance level was guided by the cognitive demands of the experiment: small sample size, equal sample groups, several cognitive tasks and multiple categories of answers. Next, the probability (p-value) that measures the evidence against the null hypothesis, i.e. the probability of either acceptance or rejection of null-hypothesis for the current empirical test is $p \leq 0.05$.

3.2. Materials and tasks

The experiment was designed in the following way: the same object by meaning (that is, by keeping its interpretant fixed) was suggested in three different ways – pictorial, verbal and diagrammatic – to experimental subjects. These three ways of object's presentation refer to Peirce's classification of signs and, in particular, to the three sign types as related to the objects of a sign – icon, index, symbol. An icon has some resemblance with the stimulus, such as something comprehended as a picture. A symbol represents by generality of its objects (such as a convention, language, text). Finally, index represents by causal connections. Hence, diagrams were chosen as the way to introduce causal connections, pictures as a way to represent iconic connections, and language/text to represent symbolic connections respectively. This typology[9] underlies the three ways in which the object was given in experimental tasks: pictorial, verbal and diagrammatic.

Each experimental task constituted a short story presented either i) pictorially (as a sequence of related pictures, e.g. comics), ii) verbally (written in language story), iii) diagrammatically (as a scheme with arrows and lines). Participants were asked to imagine the rest of the story using any method of expression. The choice of the stories was not arbitrary. Main criteria were the following: **a)** the story is easy to understand, vocabulary and formulation

[9] The same typology was applied to evaluate experimental answers as belonging to pictorial, verbal or diagrammatic categories.

of the sentences are simple and straightforward to be understood by non-native speakers with a good command of English; b) the story is concise so that the participants could easily hold in their minds the entire plot; c) the story provokes imagination, i.e. narrative and the plot that it develops is sufficiently interesting for subjects to proceed imagining the end of the story. Stories were written by actual writers, story-tellers and narrators (Chopin 2016; Baum 2016) and were chosen from material similar to children's books, ensuring the points (a)-(c). Considerable attempts were made to have all three modalities (picture, text and diagram) reflect the content of the story as precise as possible, and a professional sketch artist was used for that purpose. In total, there were three different stories presented in each of the three ways.

In addition, the choice of the stories was influenced by semantic differences and cognitive demands. It was important to choose semantically different stories, i.e. those that would put forward a different set of questions in front of the subjects and in this way would suggest different images to be produced as the solution for each of the cognitive task. The existence of such differences was expected to prove one of the sub-hypotheses of the project, namely that mental imagery is task- and context-dependent.

The expected reaction to a story-stimulus is the production of the image that is influenced by the suggested properties of the object – pictorial, verbal or diagrammatic. Thus, the final image is supposed to be different across different cognitive tasks and have distinct similar characteristics within each type of the tasks. An expected result is that the same object (by meaning) expressed in different ways produce different images/representamens.

In sum, the correlation between the image and the object is the target of the experimental investigation. Generally, each experimental task is structured in the following way: the manipulated object (story-stimulus) constitutes the independent variable, while the affected representamen (an image) constitutes the dependent variable. The fixed interpretant (i.e. the same meaning) is the control variable. Schematically, this is depicted as follows:

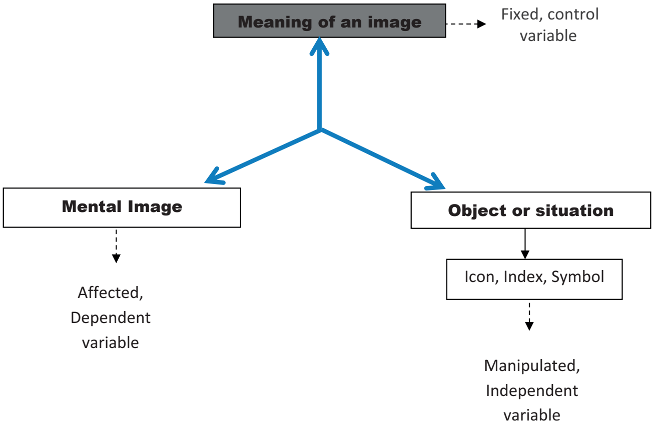


Figure II: Structure of an image experiment

3.3. Population and sampling

The target population of the experiment (i.e. the target sample) was international students of Estonia.[10] The choice of the target sample was guided mainly by the principle of convenience sampling, but with the several advances. First advantage was internationality of a sample. In current empirical study participated students of 14 different nationalities. Second advantage of sampling was differences in economic backgrounds of participants. Subjects of 20 professions or competencies took part in current study. Such diverse national, cultural and professional backgrounds of participants make them a suitable and available target population to investigate main hypothesis of the research project. Moreover, a target population of international students from various backgrounds might be a good representative of a wider international population.

At all events, the current research professes to be the first step in the investigation of the sign-theoretical account of MI. The sample of the current experimentation consisted of international students of Tallinn University of Technology (TalTech),[11] mostly bachelor-degree students. The sample size was 40 sampling units, i.e. 40 subjects participated in the experiment. Participants were chosen by the volunteer sampling. Although subjects were not randomly selected, the equal probability of participation was guaranteed by an equal information distribution among all international students via the institute's international office. A complement of free cinema tickets was offered to volunteers. Age of the participants ranged from 18 to 37, whereas an average age of the participants was 26.7 years.

Next, the experimentation was conducted among those international students whose English competence was ascertained to be very good. We asked participants' English language competence in the introductory pre-test setting. It showed that average level of language competence varied between "upper-intermediate (B2)" (42,5%) and "Advanced (C1)" (32,5%), which was taken to be sufficient for comprehension of the tasks. 10% of all participants evaluated their English language competence as "intermediate". Native English speakers constituted 15% of all participants.

All participants were randomly divided onto two groups, 20 subjects each. Each group received slightly different tasks to avoid the bias of recognizing the purpose of experimentation and to additionally test the potential differences in subjects' performance on different cognitive tasks. Instructions and task formulations were given in English and remained the same across

[10] In recent years Estonia has gained significant popularity among international students. Just for the academic year 2015/2016 Estonia has hosted 3800 international degree students, more than 1500 exchange students and ca 400 participants of summer or winter schools. (From http://www.studyinestonia.ee.)

[11] Tallinn, as being the capital of Estonia, attracts more international students compared to other cities of Estonia. So, Tallinn University of Technology currently hosts most of the international students in Estonia. For these considerations international students of TalTech were taken as a sample population of the research project.

experimental groups, but the stimulus of cognitive task differed. The number of answers was: 60 answers in each experimental group (20 subjects solved 3 tasks) and 120 answers in total. This number of answers is assumed to be large enough to show statistical relevance of the answers received and make legitimate conclusions about acceptance or rejection of the null-hypothesis.

3.4. Procedure

The experiment took place in an ordinary classroom of TalTech. In order to minimize cognitive bias and to reduce (to the extent possible) the tacit knowledge effect, the experiment was silent, i.e. participants did not know that they are participating in an empirical test, nor did they know the theoretical background or the hypotheses tested. Subjects were invited to help their university's researcher in accomplishing several game-like tasks for her doctorate dissertation. All instructions of the cognitive tasks were given directly by the experimenter before the participant started fulfilling the task. The experimenter made sure that participant understood the task and instructions by receiving a personal confirmation from the participant and answering all the questions (if there were any). During the performance of the task there was no interaction between the experimenter and the subject. The experimentation was conducted in English.

The experimentation began with a series of pilot experiments, which were conducted to check whether subjects correctly understood cognitive task, whether task instructions are clear enough, and whether the order in which tasks are given influences the responses. In total, two pilot investigations were conducted and 63 students participated in the pilot tests. Pilot experiments showed that change of the order, in which cognitive tasks are given, does not influence the responses. Additionally, subjects were sensitive towards the precise formulation of the instructions of the tasks. Thus, the results of the pilot tests helped to sharpen experimental design, formulate instructions in a clearer and more comprehensible way, eliminate the difference in task order and simplify the experiment.

The actual experiment began with the short introductory pre-test questionnaire to provide experimentally important information about participants (age, nationality, educational background, profession or field of study, English competence level). The pre-test questionnaire was followed by the three cognitive tasks given to each participant. Forty participants were divided into two groups (the study and the control groups), and were introduced with the three short stories which were the material on the three cognitive tasks for them to solve (Appendix A). The first task was a pictorial story (a sequence of pictures), second task a verbal one (text), and the third diagrammatic, including both verbal and pictorial elements. The diagram expressed the story via abstract relations (Appendix A). The order of the tasks remained the same across the two groups. By their content, these tasks were distributed in the following way: Group 1 received the first story pictorially, the second verbally and the third diagrammatically. Group 2 received the third story pictorially, the first verbally, and the second diagrammatically. Participants were equally instructed on each of the tasks as follows: "a) Look/Read carefully the story. What will happen next? b) Imagine the rest of the story, and c) Express the imagined on the next page using any method of expression". The experimentation was finished with the qualitative after-test – the Psi-Q test – where participants were invited to evaluate the "subjective vividness" of their imagery capacity.

In total, there were three stories or stimuli presented via three different modalities, totalling six tasks, with the same instructions on how to solve them. The response time was approximately 30 minutes (no sharp time constraints were given to eliminate anxiety etc.). The answers were expected to differ on various stimuli within each group and to coincide on similar stimuli across the two groups.

The experimentation was fomulated and manipulated in this way in order to be able to show that a difference in the initial traits of the imaginary object – pictorial, verbal and diagrammatic – influences "the sign" that represents this object in the final image. To minimize the tacit knowledge effect among experimental subjects and to account for cognitive biases concerning understanding the theoretical background, the content of the stories differed across three tasks within each of the group, while the stories were the same by their content across experimental groups. This ensured that "the interpretant" of the final image was fixed and repeated across experimental groups.

The after-test (Psi-Q test) was then assumed to reveal individual differences in the imaginary abilities of the subjects. That test would check for a correlation between individual imaginary capacity and the response type across three cognitive tasks. We expected those participants who estimate their scores to be high on the vividness of their MI to use a more detailed iconic imagery, while subjects with a lower vividness scores would use more symbolic or abstract imagery.

4. Results

The results of the experiment were evaluated in a categorical (nominal) scale that reflects the type of answer participants chose to produce as their final image on each of the three cognitive stimuli. There were thus three general categories: pictorial, verbal and diagrammatic. The reason for the choice of the method of classifying responses in this way[12] comes from the

[12] Noteworthy, according to Peirce's theory, there can hardly be found "pure signs", i.e. the features (symbolic+indexical, indexical+iconic etc.) are often occur to be mixed in signs. This does not preclude us to evaluate responses, following his typology of signs, as belonging to three general categories: pictorial, verbal, diagrammatic (as described

theoretical framework of Peirce's typology of signs, which corresponds to the three ways in which the objects of the stimuli are presented, namely iconic, symbolic, indexical, in the formulation of the three cognitive tasks.

The classification of the responses was made using the following reasoning: a) If the imaginary answer *resembled* its respective stimulus, then it corresponded to an iconic sign and was categorized as 'pictorial'. b) If the final image expressed *generality* (i.e. is a symbol), then it was classified as 'verbal'. c) If the produced image represented some *causal connections* (i.e. index), then it was labeled 'diagrammatic'. In this way, all three categories of answers conform to the theory as well as to the demands of the study design.

The responses of the experiment were distributed in the following way: for pictorial stimulus (Task 1) 15 answers out of 40 were given pictorially (37,5% of all respondents). For the same task, 22 answers were verbal and 3 diagrammatic (55% and 7,5%, respectively). It is noteworthy that altogether 18 answers out of 40 were given in a non-verbal way (i.e. pictorial and diagrammatic), which constituted 45% of all answers.

Table II: Answers for pictorial stimulus (Task 1).

Method/Frequency	Frequency	Percent
Diagr	3	7,5
Pictor	15	37,5
Verbal	22	55,0
Total	40	100,0

Next, for verbal stimulus (Task 2) 4 subjects out of 40 answered pictorially (10% of all respondents). For the same task we received 31 verbal answers and 5 diagrammatic (77, 5% and 12,5% of all respondents, respectively). The total number of non-verbal answers were the lowest among all three cognitive tasks, namely 9 answers (22,5% of all respondents).

Table III: Answers for verbal stimulus (Task 2).

Method/Frequency	Frequency	Percent
Diagr	5	12,5
Pictor	4	10,0
Verbal	31	77,5
Total	40	100,0

Finally, for diagrammatic stimulus (Task 3) we received 4 pictorial, 22 verbal and 14 diagrammatic answers (10%, 55% and 35% of all respondents, respectively). Similarly to the answers for Task 1, we found that the total amount of non-verbal answers was quite high: 18 answers out of 40, that is 45% of all repsondents.

above) by the most dominant/prevalent feature of the answer (i.e. either iconic, indexical or symbolic).

Table IV: Answers for diagrammatic stimulus (Task 3).

Method/Frequency	Frequency	Percent
Diagr	14	35,0
Pictor	4	10,0
Verbal	22	55,0
Total	40	100,0

No significant difference between two experimental groups and the type of the answer was found. For this reason, all results were evaluated together. The general distribution of answers across all three categories can be seen in Table IV and in the corresponding graph in Figure III:

Table V: Frequencies of answer distribution

Method/Task	**Pictorial**	**Verbal**	**Diagrammatic**
Diagrammatic	3	5	14
Pictorial	15	4	4
Verbal	22	31	22
Total	40	40	40

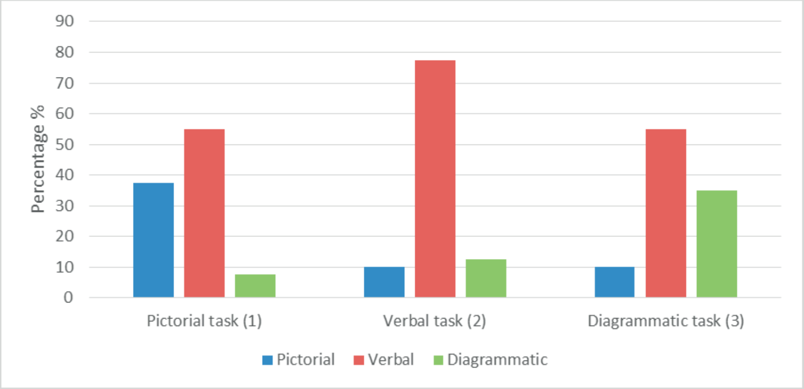


Figure III: Distribution of answers in percentage

The methods of descriptive and inferential statistics were used to analyze these results. In particular, statistical programs SPSS and R-studio were both used for statistical analysis and relevant calculations.

5. Discussion

What do these results show? Can we confirm or deny the initial hypothesis? The Pearson's Chi-squared test[13] was performed to calculate the p-value and to examine whether there is a significant relation between

[13] Pearson's Chi-squared test was chosen due to demands of current experimental design, since it allows evaluating several sets of categorical data.

properties of an object and those of an image. The R-studio statistical calculations showed $X^2(N=40) = 22{,}045$; $p = 0{,}0001963$ with df (degree of freedom) = 4. To check and confirm these calculations, we applied Fisher's test using R-studio program. The Fisher's test showed slightly different p-value, $p = 0{,}0004802$. However, the results of both tests confirmed that the relation between two variables (the method of response and the type of the task) was strongly significant, with $p <0{,}01$. Similar calculations were performed using SPSS. It confirmed previous findings with $X^2(N=40) = 22{,}045$; $p = 0{,}0001963$. Our results suggest that the null-hypothesis, namely that the characteristics of mental images remain the same regardless of the characteristics of its object, should be rejected.

The low p-value ($p \leq 0.05$) confirms the alternative hypothesis, which proposes that there is a significant interrelation between properties of an object and properties of an image. Particular properties of an object influence the characteristics of an image formed to present this object. It can be seen from Tables I-III that the distribution of answers across three types of tasks was heterogeneous. In particular, the largest number of pictorial responses (37,5%) was given on pictorial stimulus (Task 1). Similar observations hold for verbal and diagrammatic answers. Stimulus influences the formation of mental image significantly. This leads to the conclusion that mental imagery does not share certain characteristics that would be independent of the characteristics of its object. On the contrary, various properties of the object evoke various images. This challenges the idea that one cognitive format underlies the formation of mental images.

At the same time, we can observe from Tables I-III that all three response types (pictorial, verbal and diagrammatic) were used to solve the three tasks. Heterogeneity of how answers were distributed confirms the hypothesis that mental imagery cannot be understood from the perspective of one type of mental format or representation. Subjects tend to choose various methods for their image-formation that varies with multiple influencing factors. This conforms to the sign-theoretical account, as according to it there cannot be pure images of some particular type. Signs are subtle combinations of their elements and dynamic relations between them. Thus, any image might have several (that is, symbolic, iconic, indexical)[14] characteristics simultaneously. Such heterogeneity is clearly seen from the distribution of the answers in the experiment.

[14] The triadic division on icons, symbols and indices refers to Peirce's classification of signs (Peirce 1994, 1998). The explanation of the theoretically important elements of the classification is presented in another paper. In brief, a sign is an *icon* if it has a power to signify its object doe to a similarity with that object, an *index* if it has a power to signify its object due to a real relation with the object of its signification, and a *symbol* if it has a power to signify its object to an interpreter solely because it will be so interpreted. A sign can be an icon and an index simultaneously, and nothing real can be a pure icon or a pure index. Likewise, a sign can be a symbol, an icon, and an index simultaneously.

In addition, the tendency towards the mixture of answer-types was confirmed by an observational part of the study, namely the experimenter's interaction with the participants. While receiving the instructions on how to solve the tasks, several students asked whether they can use the mixture of several methods. Since this question occurred frequently, we concluded that they were inclined to use multiple modes of imagining. This could be seen as a confirmation that there is no one unified format underlying MI. Also the answers indicate this tendency towards the mixture of the response-types and the characteristics. Under a closer investigation, it occurred that students tended to use (at least) some mixture of answer-types. For example, while giving a pictorial answer to the imagined stimulus, a participant might have used arrows (diagrammatic method) to show the order of the pictures drawn; sometimes there were small linguistic 'clouds' indicating a direct speech etc. Observations did not conflict with the interpretation of the experimental data and can also be read as confirming the main hypothesis, that mental imagery, as signs, has different characteristics. Particular properties of MI are influenced by multiple factors, including characteristics of the object-stimulus, task demands, the context as well as individual differences. In brief, there is no dominant format underlying MI.

Next, we assumed that individual differences influence the image-formation. To analyze this two tests were conducted. An introductory pre-test checked whether individual variations in native language, cultural background or occupation influence mental imagery. The after-test (Psi-Q test) evaluated subjective vividness of imagery capacity and its influence on the response type. In particular, we checked whether a) participants across different backgrounds answered similarly (H0) or differently (Ha). For the Psi-Q test, we tested whether b) all participants, regardless of any subjective differences in the vividness of images, answer similarly (H0) or differently (Ha). The same programs (SPSS and R-studio) were used to statistically analyse the results. The significance level was $\alpha = 0.05$. For the first test we found no significant correlation between occupation and response type; $X^2(N=40) = 2{,}853$; $p = 0{,}415$ (with df = 3); according to Fisher's test this was $p = 0, 513$. Since the p-value was over 0.05, H_0 should be accepted. We interpret this as participants answering similarly to the three cognitive tasks independently of differences in their professions and cultural backgrounds.

Interestingly, for the second test (Psi-Q test) we found a strong correlation between individual differences in MI's vividness and response type. Analysis showed that subjects with higher vividness of MI tended to answer pictorially, producing detailed and elaborated images, whereas subjects with a lower vividness of MI tended to answer verbally, i.e. in a more abstract and general way. The significance level for this test was $p = 0{,}004$, which confirms the alternative hypothesis, namely that participants answer differently depending on subjective differences in the vividness of images. Our interpretation is that individual differences in cognitive capacities influence the formation of MI.

Taken all the above into account, a couple of general conclusions concerning the nature of mental imagery may be drawn. First, our empirical research suggests MI to exhibit characteristics that varies with multiple factors, and thus appears to be heterogeneous in nature. Our proposed interpretation is that MI would be poorly understood assuming it to be of some general or universal mental kind or format. Second, properties of mental images vary with the characteristics of the object-stimulus. MI does not share characteristics independent of the properties of an object-stimulus; rather, MI encapsulates properties of the imagined stimulus, which suggests that features of mental image depend on features of an object that it professes to represent. Finally, properties of mental images are dependent on individual differences in imaginary capacities. Indeed, human cognitive capacities surely should not be assumed to be equal: having more or less vivid imageries is well documented, both within a person and across people. Personal capacities and dispositions influence the characteristics of the produced images.

We could read these conclusions to propose that MI cannot be comprehensively explained by the prevailing representational theories that take MI to be, alternatively, matters of quasi-pictorial or propositional representations (Kosslyn 1980, 1994; Pylyshyn 1981, 2002). Our evidence showed that no dominant representational format underlies imagery. In the very least, MI can hardly be viewed as a static mental representation of a fixed particular format, which is implied by computational-repesentational paradigm (Clapin 2002; Marr 2006). Rather images change their characteristics dependent on the context, task, and the features of the objects. Dependence of an image-formation on the characteristics of its co-related elements strongly suggests that MI is not the matter of a static representation, and that dynamic mental activity occurs within the context of the creation of mentally depicted relations.

A coherent account on the nature of MI would explain such features as the heterogeneity of its characteristics, its task-context-object dependence, and the influence differences have on the image formation, among others. It might be difficult to explain all these facts by traditional representational theories of MI. Although, quasi-pictorial theory could easily accommodate pictorial data, whereas propositional theory – verbal data, the explanation of the current results by traditional accounts will still remain partial. The restriction of MI to quasi-pictorial – propositional dichotomy inevitably neglects at least some of the above-stated characteristics of MI. The reason for this might be hidden in the implicit demand of the dominant computational paradigm: the search of the primary code, which would unravel the complex mechanisms of human mind. However, new research methods and empirical data show that above-stated demand might be misleading (von Eckardt 1993; Bechtel 1998; Knuutila 2005, 2011). The results of current empirical study confirm this idea. Similarly, enactivist theory can potentially explain the dynamic relations and task-object dependence, but it could hardly account for divergent characteristics of images. Enactivism lacks a comprehensive explanation of MI's structure and diversity.

In contrast, the sign-theoretical approach that we have advocated can accommodate heterogeneity of MI's properties, its task-context-object dependence, and individual differences in imaginary capacities within one framework. First, MI can be seen as a sign that consists of three elements: representamen, object and interpretant. Taking MI to be of this triadic structure allows a detailed explanation of the nature and function of images in human cognition. Second, the theory of signs proceeds to take mental capacities to be of signifying nature. This would connect MI with many other cognitive abilities of the human mind, and would explain individual differences and dispositions in the creation of MI. Third, the theory is concerned with the dynamic and open nature of semiotic relations between the three elements of a sign. This allows it to be applied to the explanation of divergent and changing properties of mental images.

Although experimental findings support the theory of sign towards investigation of MI, one might also argue that there are certain weaknesses in experimental studies conducted on MI. First, our three cognitive tasks might evoke different cognitive capacities (e.g., decision-making, creative thinking etc.) as well. How can one be sure that it was MI that was used to solve these experimental tasks? Now the employment of mental images was ensured by the precise and detailed instructions given by the experimenter and by receiving personal confirmation that each participant understood the task. Nevertheless, it is not straightforward to separate MI from other cognitive faculties or to eliminate their influence on the response rate. Also, the design of the task is not free from criticism, especially if compared with previous studies on similar matters. To this we reply as follows. In contrast to standard cognitive tasks testing MI where subjects are asked to memorize some stimulus, our task to *imagine* the rest of the story is markedly different. It allows testing the production of an image in natural way, such as what people might use in their daily life while planning, thinking, analysing, reasoning or daydreaming. There are in fact indications that two-thirds of our waking life mind is actually wandering and not well controlled or self-controlled by us, the self, by some cognitive agency (Metzinger 2017). In light of mind-wandering theories, it is only natural to test the nature of MI in the proposed manner. The MI produced as a response to our cognitive task is not artificial but spontaneous and may in fact be more 'ecologically valid' – subjects were free to choose any method to imagine and were not asked or expected in any way to remember the stimulus.

Further, one could read the results of the study differently, saying that shown heterogeneity of imaginary characteristics might be understood on phenomenal level only, while internal structure of MI remains one and the same. In this case, internal structure of MI is supposed to be hidden and consciously[15] inaccessible. While current study does not apply brain-scanning

[15] According to the sign-theoretic account, there are several levels of conscious access to a mental sign (for details, see Champagne 2018). The 'sub-personal' level corresponds to single sensations and qualities that are registered by the mind, but are not yet attended by

techniques in demonstrating the underlying difference in neurophysiological terms, it does show the structural difference of the various images produced as a response on different stimuli, which implies the former. If the representational format of MI would indeed be sub-personally and unconsciously one and the same across different images, then the answers to imaginary tasks would be expected to be similar. But this is not the case. Moreover, one would expect one and the same subject to answer similarly to all stimuli, but this was not the case either. One and the same subject typically used various types of images to answer different stimuli. Based on the above-stated data this difference is statistically significant. Thus, we are inclined to conclude that shown heterogeneity of MI's properties is not (just) phenomenal: the difference in the modes of expression of an internal image does say something about sub-personal and unconscious level of image formation.

Finally, the limited sample size and the volunteer sampling method instead of full randomization may be a limitation of the current research, conducted under limited organisational allowances, and random sampling method and larger sample size is suggested for replication.

To sum up, limitations notwithstanding, explanations of MI should not overlook the potential of seeing them as *signs*. A sign-theoretical approach might overcome some long-standing controversies and limits of the prevailing representational accounts. In particular, the experimental approach suggests a new perspective where divergent properties of mental images come together under the umbrella of a Peircean theory.

5. Conclusion

We studied the nature of mental imagery by an experiment in the theoretical context of Peirce's semiotics (the theory of signs). An empirical test was carried out that hypothesised that mental imagery can be accounted for in that theory. According to the theory, MI is a sign that consists of three relata: representamen, object and interpretant, and it is characterized by dynamic and context-dependent semiotic relationships. To test the hypothesis, an experiment was designed. The analysis of the results showed that 1) the characteristics of mental images are heterogeneous in nature; 2) properties of mental images are dependent on the characteristics of object-stimulus; 3) properties of mental images are dependent on individual differences in imaginary capacities. These results were interpreted to indicate that, contrary to standard representational accounts, MI does not emerge from one dominant

the conscious self. The internal structure of the mind and MI, however, does not change dependent on the level of conscious accessibility. Higher level of conscious access – which is the level of current study – is the indicator of the internal (sub-personal/hidden) structure. The sign-theoretic account that we are applying can accommodate these levels within one theoretical framework.

representational format (such as quasi-pictorial or propositional). Standard representational accounts may fail to provide comprehensive explanations of heterogeneous characteristics of MI and their context dependence. Our study concludes that these features can, however, be explained by Peirce's theory of signs. The results support the idea that MI can be seen as signs. Under that light, MI constitute complex mental phenomena with manifold traits and dynamic, continuously changing relations between its elements. While new empirical investigations that exploit the sign-theoretical approach are needed, this interpretation of the results of the present experiment is also a strong indication that the theory of signs is a viable methodological alternative that accommodates heterogeneous empirical evidence.

Acknowledgments:

This empirical research was founded by Estonian Research Council scientific grant PUT1305 "Abduction in the Age of Fundamental Uncertainty", 2017. We would like to thank Amirouche Moktefi (PhD) for his valuable advices and constant help, as well as Liisa Kruusamägi – a talented illustrator – for her work and patience. We would like to express a special gratitude to Kristiina Averin – philosopher and psychologist – for her indispensable help with statistical analysis. The authors declare no conflict to interest.

References:

Andrade, Jackie, May, Jon, Deeprose, Catherine, Sarah-Jane Baugh and Giorgio Ganis. 2013. Psi-Q: the Plymouth Sensory Imagery Questionnaire. *British Journal of Psychology* 105. 547–563.

Anderson, John R. 1978. Arguments concerning representations for mental imagery. *Psychological Review* 85. 249–277.

Atkin, Albert. 2013. Peirce's theory of signs. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/sum2013/entries/peirce-semiotics/. (accessed 15 September 2017).

Bartolomeo, Paolo 2008. The neural correlates of visual mental imagery: an ongoing debate. *Cortex* 44(2). 107–108.

Baum, Stuart B. 2017. The blue bottle. http://www.stuartstories.com/activities/bluebottle.html. (accessed 15 September 2017).

Bechtel, William. 1998. Representations and cognitive explanations: assesing the dynamicist's challenge in cognitive science. *Cognitive Science* 22(3). 295–318.

Betts, George Herbert. 1909. The distribution and functions of mental imagery. *Teachers' College Columbia University Contributions to Education* 26. 1–99.

Chambers, Deborah, Reisberg, Daniel. 1985. Can mental images be ambiguous? *Journal of Experimental Psychology: Human Perception and Performance* 11. 317–328.

Champagne, Mark. 2018. *Consciousness and the Philosophy of Signs. How Peircean Semiotics Combines Phenomenal Qualia and Practical Effects*. Dordrecht: Springer.

Chopin, Kate. 2017. The pair of silk stockings. http://www.eastoftheweb.com/short-stories/UBooks/PairSilk859.shtml. (accessed 15 September 2017).

Clapin, Hugh. 2002. *Philosophy of mental representation*. Oxford: Clarendon Press.

Dulin, David, Hatwell, Yvette, Pylyshyn, Zenon W., Chokron, Sylvie. 2008. Effects of peripheral and central visual impairment on mental imagery capacity. *Neuroscience and Biobehavioral Reviews* 32. 1396–1408.

Finke, Ronald A., Pinker Steven, Farah, Martha J. 1989. Reinterpreting visual patterns in mental imagery. *Cognitive Science* 13. 51–78.

Fodor, Jerry A. 1975. *The language of thought*. Cambridge, Mass.: Harvard University Press.

Fodor, Jerry A. 1990. *A theory of content and other essays*. Cambridge, Mass.: MIT Press

Ganis, Giorgio. 2013. Visual mental imagery. In Simon Lacey and Rebecca Lawson (eds.), *Multisensory imagery,* 9–28. Dordrecht: Springer.

Gordon, Rosemary. 1949. Investigation into some of the factors that favour the formation of stereotype images. *British Journal of Psychology* 39. 156–167.

Gregg, Melanie J. & Clark, Terry. 2007. Theoretical and practical applications of mental imagery. In Aaron Williamon and Werner Goebl (eds.), *Proceedings of the international symposium on performance science 2013*, 295–300. Brussels, Belgium: European Association of Conservatoires (AEC).

Issayeva, Jelena. 2015. Sign theory at work: The mental imagery debate revisited. *Sign Systems Studies* 43(4), 584–596.

Keller, Peter E. 2012. Mental imagery in music performance: underlying mechanisms and potential benefits. *Annals of the New York Academy of Sciences. The neurosciences and music IV learning and memory 1252*. 206–213.

Knuuttila, Tarja. 2005. *Models as epistemic artefacts: toward a non-representationalist account of scientific representation.* Dissertation. University of Helsinki.

Knuuttila, Tarja. 2011. Modelling and representing: an artefactual approach to model-based representation. *Studies in History and Philosophy of Science* 42. 262–271.

Kosslyn, Stephen M. 1978. Measuring the visual angle of the mind's eye. *Cognitive Psychology* 10. 356–89.

Kosslyn, Stephen M. 1980. *Image and mind*. Cambridge, Mass.: Harvard University Press.

Kosslyn, Stephen M. 1988. Aspects of a cognitive neuroscience of mental imagery. *Science* 240(4859). 1621–1626.

Kosslyn, Stephen M. 1994. *Image and brain: the resolution of the imagery debate*. Cambridge, Mass.: MIT Press.

Kosslyn, Stephen M., Ganis, Giorgio, Thompson, William L. 2006. Mental imagery and the human brain. In Qicheng Jing, Mark R. Rosenzweig, Gerry d'Ydewalle, Houcan Zhang, Hsuan-Chih Chen, Kan Zhang (eds.), *Progress in psychological science around the world: neural, cognitive and developmental issues,* 195–209. New York, NY: Psychology Press.

Lacey, Simon & Lawson, Rebecca. 2013. *Multisensory imagery*. Dordrecht: Springer.

Marks, David F. 1973. Visual imagery in the recall of pictures. *British Journal of Psychology* 64. 17–24.

Marr, David. 2006. Vision. In Jose L. Bermudez (ed.), *Philosophy of psychology. Contemporary readings,* 385–406. New York and London: Routledge.

Merrell, Floyd. 2001. Charles Sanders Peirce's concept of the sign. In Paul Cobley (ed.), *The Routledge companion to semiotics and linguistics*, 28–39. New York and London: Routledge.

Metzinger, Thomas. 2017. Why is mind-wandering interesting for philosophers? In Kieran C.R. Fox & Kalina Christoff (eds.). *The Oxford Handbook of Spontaneous Thought: Mind-wandering, Creativity, Dreaming, and Clinical Conditions*. New York, NY: Oxford University Press.

Millikan, Ruth G. 1984. *Language, thought and other biological categories*. Cambridge, Mass.: MIT Press.

Moro, Valentina, Berlucchi, Giovanni, Lerch, Jason, Tomaiuolo, Francesco, Aglioti, Salvatore M. 2008. Selective deficit of mental visual imagery with intact primary visual cortex and visual perception. *Cortex* 44. 109–118.

Morris, Charles W. 1938. *Foundations of the theory of signs.* Chicago, IL: The University of Chicago Press.

Morris, Charles W. 1946. *Signs, language and behavior*. New York, NY: Prentice-Hall.

Morris, Charles W. 1964. *Signification and significance: a study of the relations of signs and values.* Cambridge, Mass.: MIT Press.

Pascual-Leone, Alvaro, Nguyet, Dang, Cohen, L.eonardo G., Brasil-Neto, Joaquim P., Cammarota, Angel, Hallett, Mark. 1995. Modulation of muscle responses evoked by transcranial magnetic stimulation during the acquisition of new fine motor skills. *Journal of Neurophysiology* 74(3). 1037–45.

Paivio, Allan U. 1971. *Imagery and Verbal Processes*. New York, NY: Holt, Rinehart and Winston.

Paivio, Allan U. 1986. *Mental Representations: A Dual Coding Approach*. New York, NY: Oxford University Press.

Pearson, Joel & Kosslyn, Stephen M. 2015. The heterogeneity of mental representation: ending the imagery debate. *Proceedings of the National Academy of Sciences of United States of America* 112(33). 10089–10092.

Peirce, Charles S. 1994. *The collected papers of Charles Sanders Peirce*. Charles Hartshorne and Paul Weiss (eds.). Cambridge, Mass.: Harvard University Press. Vols. VII-VIII (1958), Arthur W. Burks (ed.), Cambridge, Mass.: Harvard University Press.

Peirce, Charles S. 1998. *The essential Peirce. Selected philosophical writings. Volume 2 (1893–1913)*. Nathan Houser, Jonathan R. Eller, Albert C. Lewis, Andre D. Tienne, Cathy L. Clark, D. Bront Davis (eds.). Bloomington and Indianapolis: Indiana University Press.

Perky, Mary C. W. 1910. An experimental study of imagination. *American Journal of Psychology* (21). 422–52.

Pietarinen, Ahti-Veikko. 2012. Peirce and the logic of image. *Semiotica* (192), 251–261.

Pietarinen, Ahti-Veikko. 2015. Signs systematically studied: Invitation to Peirce's theory. *Sign Systems Studies* 43(4), 372–398.

Pinker, Steven & Finke, Ronald A. 1980. Emergent two-dimensional patterns in images rotated in depth. *Journal of Experimental Psychology: Human Perception and Performance* (4). 21–35.

Plessinger, Annie. 2007. The effects of mental imagery on athletic performance. http://healthpsych.psy.vanderbilt.edu/HealthPsych/mentalimagery.html (accessed 15 September 2017).

Pylyshyn, Zenon W. 1973. What the mind's eye tells the mind's brain: the critique of mental imagery. *Psychological Bulletin 80*(1). 1–24.

Pylyshyn, Zenon W. 1981. The imagery debate: analogue media versus tacit knowledge. *Psychological Review* 88. 16–45.

Pylyshyn Zenon W. 2000. Situating vision in the world. *Trends in Cognitive Sciences* 4(5). 197–206.

Pylyshyn, Zenon W. 2002. Mental imagery: in search of a theory. *Behavioral and Brain Sciences* 25(2). 157–238.

Pylyshyn, Zenon W. 2003. *Seeing and visualizing: it's not what you think*. Cambridge, Mass.: The MIT Press.

Richardson, Peggy A. 1995. Therapeutic imagery and athletic injuries. *Journal of Athletic Training* 30(1). 10–12.

Rock, Irvin, Wheeler, Deborah & Tudor, Leslie. 1989. Can we imagine how objects look from other viewpoints? *Cognitive Psychology* 21(2). 185–210.

Saussure, Ferdinand. 1983. *Course in general linguistics*. Translated by Roy Harris. London: Duckworth.

Schmidt, Timo T., Ostwald, Dirk, Blankenburg, Felix. 2014. Imaging tactile imagery: changes in the brain connectivity support perceptual grounding of mental images in primary sensory cortices. *Neuroimage* 98. 216–24.

Sheehan, Peter W. 1967. A shortened form of Betts' questionnaire upon mental imagery. *Journal of Clinical Psychology* 23. 386–389.

Shepard, Roger N. & Metzler, Jacqueline. 1971. Mental rotation of three-dimensional objects. *Science* 171. 701–703.

Shepard, Roger N. & Cooper, Lynn A. 1982. *Mental images and their transformations*. Cambridge, Mass.: MIT Press.

Sima, Jan Frederik. 2011. The nature of mental images – an integrative computational theory. In Laura Carlson, Christoph Hölscher, Thomas F. Shipley (eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society,* 2878–2883. Austin: TX.

Slezak, Peter. 1990. Re-interpreting images. *Analysis* 50(4). 231–243.

Slezak, Peter. 1991. Can images be rotated and inspected? A test of the pictorial medium theory. *The Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. 55–60.

Slezak, Peter. 1995. The 'philosophical' case against visual imagery. In Peter. Slezak, Terry Caelli, Richard Clark (eds.), *Perspectives on cognitive science: theories, experiments and foundations,* 237–271. Norwood, NJ: Ablex Publishing.

Slotnick, Scott D., Thompson, William L., Kosslyn, Stephen M. 2005. Visual mental imagery induces retinotopically organized activation of early visual areas. *Cerebral Cortex* 15(10). 1570–1583.

Thomas, Nigel J. T. 2009. Visual imagery and consciousness. In William P. Banks (ed.), *Encyclopedia of consciousness*, 445–457. Oxford: Academic Press/Elsevier.

Thomas, Nigel J. T. 2010. Imagery and coherence of imagination. *Journal of Philosophical Research* 22. 95–127.

Thomas, Nigel J. T. 2014. Mental imagery. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy.* http://plato.stanford.edu/archives/fall2014/entries/mental-imagery/. (accessed 15 September 2017).

Tye, Michael. 1991. *The imagery debate*. Cambridge, Mass.: MIT Press.

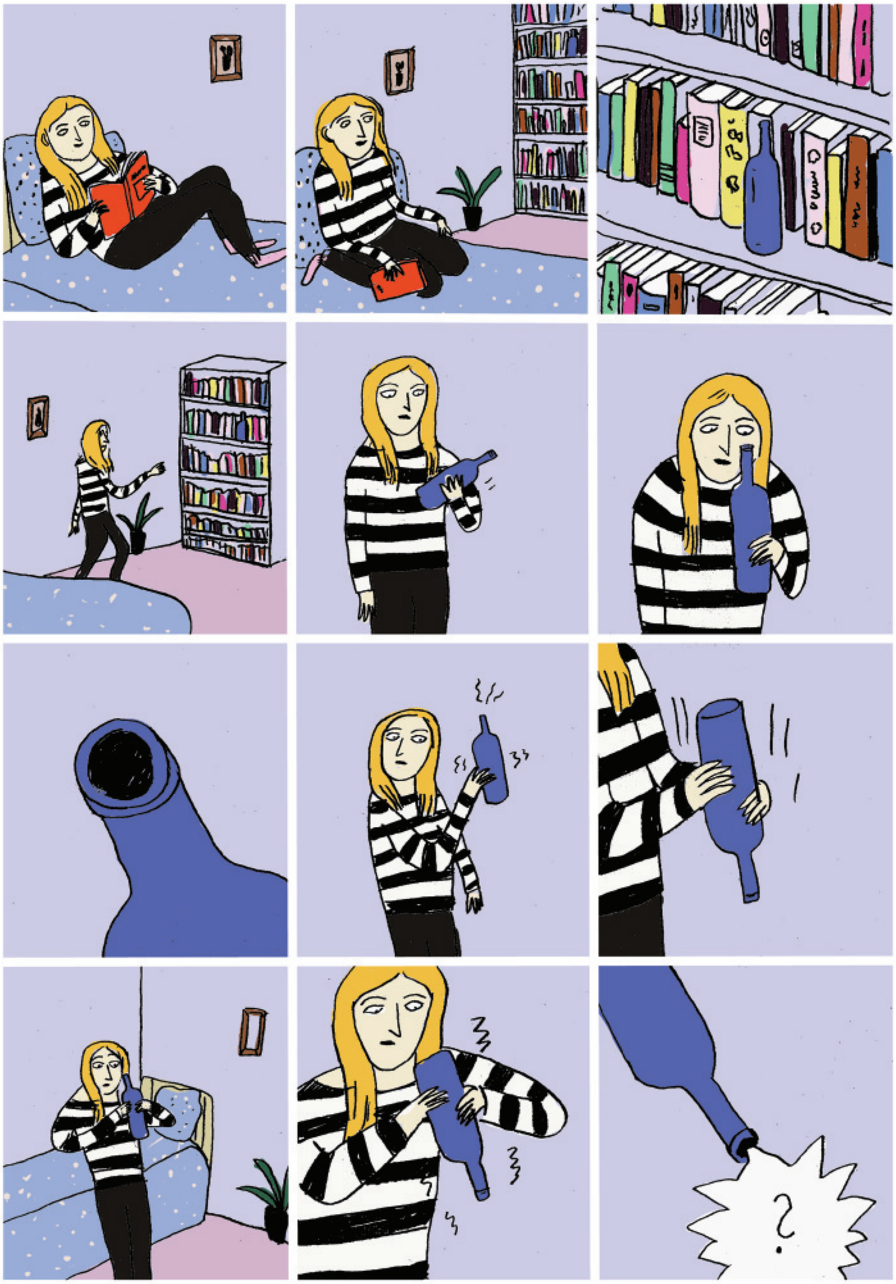
Van Gelder, Tim. 1995 What might cognition be, if not computation? *The Journal of Philosophy* 92. 345–381.

Von Eckardt, Barbara. 1993. *What is cognitive science*? Cambridge, Mass.: MIT Press.

Zeman, Jay. 1977. Peirce's theory of signs. In Thomas A. Sebeok (ed.), *A Perfusion of signs*, 22–39. Bloomington: Indiana University Press.

Appendix A: The Examples of cognitive tasks

Task 1: a) Look carefully at the story. What will happen next? **b)** Imagine the rest of the story, and **c)** Express the imagined on the next page using any method of expression.

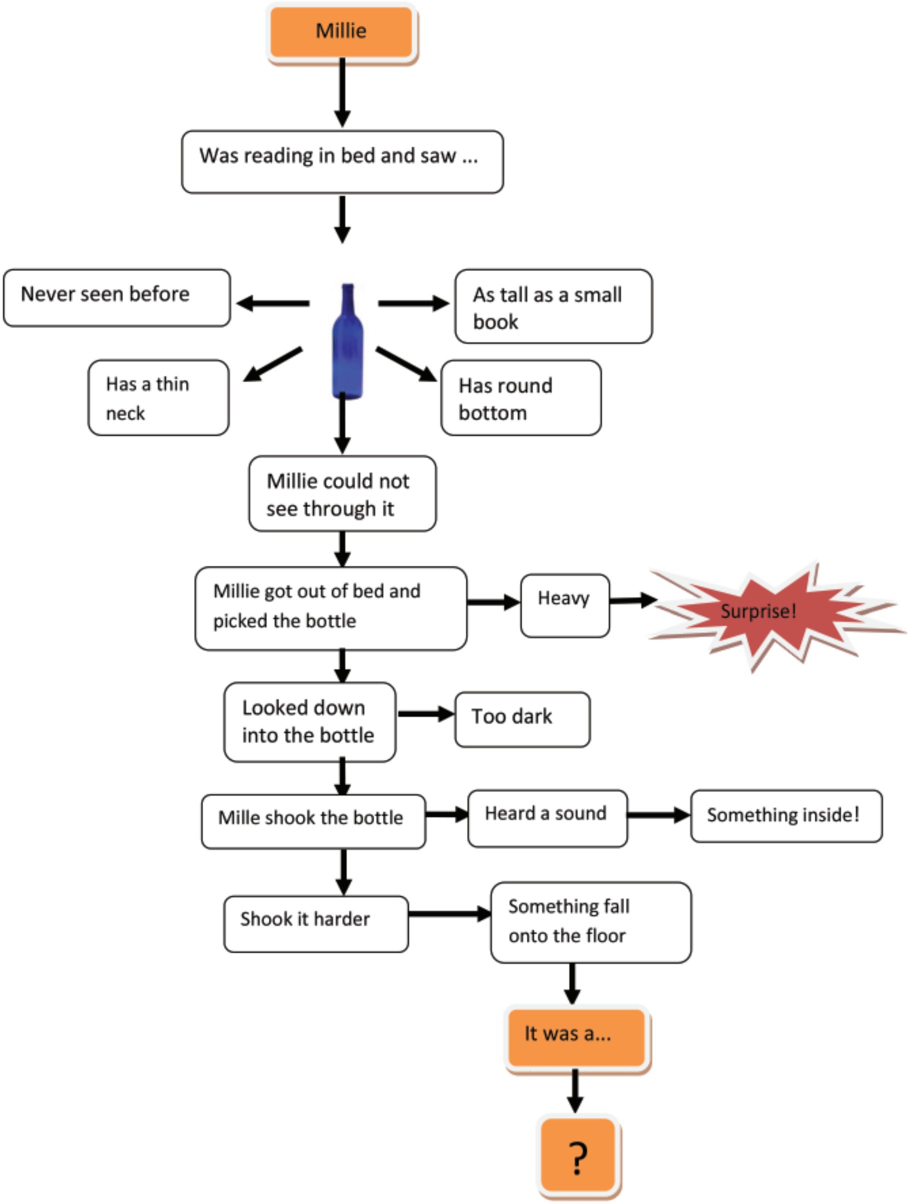


Task 2: a) Read carefully the story. What will happen next? **b)** Imagine the rest of the story, and **c)** Express the imagined on the next page using any method of expression.

One night, Millie was up late reading in bed. She finished the book she was reading and looked over to her shelves to see what else she might read before she went to bed. Right there on her shelf was something she had never seen before. It was a blue bottle. The blue bottle was about as tall as a small book, had a round bottom, and a thin neck. And while the bottle looked as if it were made out of glass, Millie could not see through it. Millie got out of bed and went over to the bottle. She picked it up, carefully, afraid that it might break. She was surprised at just how heavy it was. Certainly heavier than any other bottle this size she had ever before lifted.

She looked down into the bottle, but it was too dark inside to see anything. So she shook it. She heard a rattling sound. There was something inside! She turned the bottle upside down and shook it again, to see if anything would fall out. Something almost fell out and then it didn't. Whatever was inside was now stuck in the bottle's neck. Millie shook harder and harder. Finally, something small fell onto the floor. It was a ...

Task 3: a) Look carefully at the story. What will happen next? **b)** Imagine the rest of the story, and **c)** Express the imagined on the next page using any method of expression.



William Ramsey
Department of Philosophy
University of Nevada, Las Vegas

Original Scientific Paper
UDK 159.956:165.1

THE COGNITIVE SOURCE OF PHILOSOPHICAL INTUITIONS: TROUBLES FOR THE ETIOLOGICAL PROJECT*

Abstract: *The advent of experimental philosophy has generated a renewed interest in the nature of philosophical intuitions. This has led many to assume that understanding the psychological processes that generate philosophical intuitions will provide a much needed source of answers for various questions about their nature. It is widely assumed that if we are to gain a good sense of the reliability of philosophical intuitions (and the degree to which experimental philosophy challenges their role in philosophy) it is critical to learn more about the sort of cognitive faculties and processes that produce them. However, here, I'm going to present a more pessimistic outlook about what we should expect from such a project in the case of philosophical intuitions. More specifically, I'll suggest that there are two central problems that prevent a deeper understanding of the cognitive source of intuitions from doing much of what we would like it to do. The first problem is that for an important range of philosophical intuitions, it is doubtful that there is a stable set of cognitive processes and mechanisms generating them that is sufficiently universal to allow for the sort of generalizations we would like. Philosophical intuitions are likely created by an assortment of varying psychological operations, and that learning about how a particular intuition is created is unlikely to tell us much about how other intuitions are formed. In short, there is no faculty of intuition in any meaningful sense. I call this the "Etiological-Diversity Problem". The second problem stems from the fact that one of the primary things we want knowledge of the intuition-producing cognitive machinery to tell us about is the reliability of the intuitions they produce. However, this expectation gets the normal order backwards. The evaluation of cognitive processes or mechanisms typically requires a prior evaluation of the normative status of the states it produces in different contexts, distinguishing proper functioning and competence from improper functioning and performance errors by virtue of the quality of the mechanism's output. It is far less clear how this order can be reversed. Consequently, it is hard to see how an understanding of intuition-producing mechanisms will allow us to evaluate the epistemic status of philosophical intuitions without already having such an evaluation in hand. I call this the "The Secondary Calibration Problem". The upshot is that even though I believe there is much to learn about the etiology of intuitions that will be useful, these problems severely hamper our ability to answer many of the central questions we have about the nature and trustworthiness of philosophical intuitions.*

Keywords: *etiological project, cognitive program, etiological diversity, calibration problem*

* Work on this paper was supported in part during sabbatical leave funded by the University of Nevada, Las Vegas. I am grateful to two anonymous reviewers for their extremely helpful comments and suggestions, and also to Miljana Milojevic for her valuable and insightful recommendations.

doi: 10.5937/BPA1831085R

I. Introduction

The discoveries and theoretical developments associated with experimental philosophy have turned new attention to philosophical intuitions – upon their nature, their reliability, their stability, and a host of other dimensions. With this recent interest, it has become increasingly common for authors to suggest that understanding the psychological processes that generate philosophical intuitions will provide a much needed source of answers for various questions about their nature. More specifically, it is generally assumed that if we are to gain a good sense of the reliability of philosophical intuitions (and the degree to which experimental philosophy challenges their role in philosophy) it is critical to get a firmer grasp of the sort of cognitive faculties that produce them. For example, Helen De Cruz states that ". . . understanding the etiology of philosophical intuitions is vital for making headway in debates on their evidential value" (De Cruz, 2015, p. 235). Similarly, Brian Talbot tells us that the best way to address concerns about philosophical intuitions is to "develop a general and systematic understanding of how intuitions work: where they come from, how they are generated, what they are and are not based on, what factors affect them" (Talbot, 2009, p. 159). Echoing these themes, Joshua Alexander asserts that, "[w]hat is needed, then, is a better understanding of the cognitive processes involved in the production of our philosophical intuitions. By coming to better understand what intuitions are, where they come from, and what factors influence them, we can better understand what role they can play in philosophical practice" (Alexander, 2012, p. 3).[1]

For the most part, I am sympathetic with the idea that a good way to enrich our understanding of some psychological state or condition is to attend to its etiology. If we want to know the nature of some mental state and its relation to the world, then it certainly makes sense to attend to how such states are formed – to the cognitive machinery and conditions that generate them. However, despite the prima facie reasonableness of many such projects in psychology, I'm going to present a more pessimistic outlook about what we should expect from such a project in the case of philosophical intuitions. More specifically, I'll suggest that there are two central problems that prevent a deeper understanding of the cognitive source of intuitions from doing much of what we would like it to do. The first problem is that for an important range of philosophical intuitions, it is doubtful that there is a stable set of cognitive processes and mechanisms generating them that is sufficiently universal to allow for the sort of generalizations we would like. That is, for an important class of philosophical intuitions, they are likely created by an assortment of varying psychological operations, and that learning about how a particular intuition is created is unlikely to tell us much about how other intuitions are formed. In short, there is no faculty of intuition in any meaningful sense. I

[1] For further examples of this perspective, see Allman and Woodward, 2008; Fischer and Collins, 2015; Knobe and Nichols, 2008; Knobe and Prinze, 2008; Scholl, 2007.

call this the "Etiological-Diversity Problem". The second problem stems from the fact that one of the primary things we want knowledge of the intuition-producing cognitive machinery to tell us about is the reliability of the intuitions they produce. However, this expectation gets the normal order backwards. The evaluation of cognitive processes or mechanisms typically requires a prior evaluation of the normative status of the states they produce in different contexts. We typically distinguish proper functioning and competence from improper functioning and performance errors by virtue of the quality of the mechanism's *output*. It is far less clear how this order can be reversed – how we can use understanding of the functioning of a cognitive process to then evaluate the nature of the cognitive state it generates. Consequently, it is hard to see how an understanding of intuition-producing mechanisms will allow us to evaluate the epistemic status of philosophical intuitions without already having such an evaluation in hand. I call this the "The Secondary Calibration Problem". The upshot is that even though I believe there is much to learn about the etiology of intuitions that will be useful, these problems severely hamper our ability to answer many of the central questions we have about the nature and trustworthiness of philosophical intuitions.

To show all this, the next section will fill in some of the background on what I will call the "etiological project"[2], briefly looking at some of the programs of cognitive research where it has been pursued in understanding the way philosophical intuitions are generated. Then, in Section III, I will present the etiological-diversity problem and explain how it is likely to undermine our ability to make significant generalizations about the way philosophical intuitions are produced. In Section IV, I will present the secondary calibration problem and explain how it poses a barrier to our epistemic evaluation of intuitions. Section V offers a brief conclusion and some recommendations about how to move forward in light of these considerations.

II. Investigating the Etiology of Intuitions: The Why and the How

The term "philosophical intuition" is widely recognized as having an unclear extension and refers to a range of different types of states and attitudes (Nado, 2013). To focus our discussion, I will restrict my analysis to the sort of intuitions generated by the consideration of a philosophical vignette or thought experiment, the scenarios that comprise what is often referred to as "the method of cases". Famous examples commonly discussed in the literature include the intuitive judgments prompted by Gettier's 10 Coin Case, Kripke's Gödel cases, Thompson's violinist case, and so on. We can refer to this class of intuitions as Thought Experiment Intuitive Judgments[3]

[2] This has also been referred to as "the sources project" (Pust, 2017) and the "cognitive program" (Sytsma and Livengood, 2016).

[3] The term 'judgment' is ambiguous, in that it refers to both a process (the judging) and the end result of the process (the resulting judgment). Here I am using the term in the

or "TEIJ"s for short. Even with this restriction, there are still debates about the psychological nature of such states. Some regard them as a type of belief (e.g., Lewis, 1983), while others suggest they are an inclination or disposition to believe (Sosa, 1998; Earlenbaugh and Molyneux, 2009), while still others argue that they are a distinctive type of "intellectual seeming" or a distinct type of propositional attitude (Bealer, 1998). The latter strikes me as the most plausible perspective, although not a lot of what I will say will depend upon this. Writers have also emphasized the distinction between the psychological state – the intuiting – and the content of such a state – the intuited (Lycan, 1988). Our discussion here will focus upon the psychological processes and mechanisms that produce the former.

Let's begin by stepping back and considering why we want to study the etiology of intuitions in the first place, and then we can consider how such a study is likely to go. As suggested by the quotations in the introduction, the desire to learn about the cognitive processes and mechanisms that create philosophical intuitions is grounded in the assumption that doing so will help us answer a number of important questions about their nature. A couple of central questions we assume studying the etiology of intuitions will help us answer are these:

1. Are the cognitive processes and mechanisms that produce intuitions trustworthy – do they produce accurate representations of facts concerning philosophical matters, and if so, how?
2. Insofar as such mechanisms occasionally go awry, what sorts of conditions and situation are likely to lead to their improper operation and thus to the production of faulty intuitions, and how can these conditions and situations be avoided?

The linking of these sorts of questions and empirical work on the etiology of intuitions is, at least initially, easy to understand. If you want to know about the reliability of a cognitive mechanism, then you need to understand its actual nature – its basic operation, the sub-systems involved, the sort of computational processes involved, its responsiveness to mind-independent reality, and so on. If you want to know how it properly functions, and to be able to distinguish proper function from malfunction, then, at the very least, you need to learn about its normal operations. Moreover, learning about the functioning of a faculty is often a critical step in learning about the nature of the sort of cognitive state it generates.

There are couple of important things to notice about these questions. The first is their clearly normative nature. Notions like trustworthiness or proper and improper operation demand more than a straightforward mechanical description of the system's operations; they demand that we make value assessments that rank some types of functioning as superior to

latter sense – referring to the state produced, the relevant intellectual seeming that such and such is the case.

others. We want to gain knowledge about the processes and mechanisms that create TEIJs because we think such knowledge can tell us something about the *quality* of those intuitive states. As we will see below, establishing this normative dimension for the cognitive machinery responsible for an important class of intuitions is going to be deeply problematic. The second thing to note is that questions like these are related; answers to some clearly depend upon the answers to others. Questions about malfunction or about flawed performance presuppose that we can ascertain when a faculty is functioning properly, and that clearly requires knowledge about the process whereby it generates appropriate outputs. To some degree, questions like these stand or fall together. Difficulties with answering any one of them will likely raise significant problems for answering others.

So much for the why – what about the how? How should the etiology of intuitions be investigated? While some have suggested that intuitions can be adequately investigated using introspection or some sort of a priori reasoning, there is a growing consensus that the production of philosophical intuitions is an empirical matter and should be investigated by the relevant areas of cognitive science (Talbot, 2009). Like any other sort of cognitive state, philosophical intuitive judgments are first and foremost types of psychological phenomena that should be treated as such. Still, this constraint allows for a very broad range of possible research programs and approaches. Pertinent research has come from (and likely will continue to come from) an array of different research programs, including everything from low-level neurological studies and imaging to much more abstract theorizing about computational processes. Indeed, just a casual survey reveals several different research approaches that have yielded accounts of the cognitive processes responsible for intuitions. While these approaches are not completely distinct and often overlap a great deal, they nevertheless involve their own presuppositions, conceptual apparatus and over-arching perspectives. Three prominent research approaches to understanding the cognitive source of TEIJs are: 1) the concept application/categorization approach, 2) the dual-process framework and, 3) the heuristics approach. It will help to briefly consider each of these.

A. The Concept Application/Categorization Framework

It is often assumed that for TEIJs, the judgment stems from the application of a philosophical concept to a particular situation. The content of nearly all TEIJs can be characterized as having the form: 'the X in this case is an instance of Y' (e.g., the belief in this case is an instance of non-knowledge; the act in the case in question is an instance of immorality; the behavior in this case is an instance of free action, and so on). As Goldman puts it, intuitions "are what might be called *classification* or *application* intuitions, because they are intuitions about how cases are to be classified, or whether various categories or concepts apply to selected cases" (2007, p. 4). Thus, one clear area of research used to understand the production of TEIJs is psychological research on concept application. Within this thriving area of research, there have been three central theoretical paradigms about concepts and how they are employed in categorization judgments: the prototype paradigm, the exemplar paradigm and the theory-based account. Each of these theoretical frameworks has demonstrated success in explaining important findings with regard to the way we categorize entities, kinds, properties, events, and the various other things concepts are about. For example, within the prototype paradigm, categorization is usually driven by matching features possessed by the target with weighted features that comprise the concept. This approach has had considerable success in explaining what has come to be referred to as "typicality effects" (Rosch, 1978), such as our tendency to rank members of a kind as more typical than others, (Smith and Medin, 1981). By contrast, with the exemplar tradition concepts are stored familiar instances encountered (Barsalou, L. W., Huttenlocher, J., & Lamberts, K., 1998), so that, say, one's concept of DOG would be represented by a familiar dog stored in memory, like the family pet Fido. This helps explain the speed with which we categorize atypical but familiar instances. The third tradition – the theory-based account – claims that concepts are represented by tacit theories about the concept's target (Keil, 1989). This model explains the various ways in which our categorization activity looks more like the application of a theory rather than a comparison to some internally stored representation.

On the concept application approach, when imagining the scenario presented by a philosophical thought experiment, we classify something or someone in the scenario as qualifying (or failing to qualify) as an instance of a philosophical property by making a categorization judgment using something like a prototype, exemplar or tacit theory. We will look more closely at how these three models of concepts could be employed to explain philosophical intuitions in the next section. As a preview, there is good reason for thinking that all three of these models of concept application are employed by our brains. Thus, even if we restrict the etiological project to the concept application/categorization approach, it is likely that the way in which we form TEIJs varies dramatically between different mechanisms and operations.

B. The Dual Process Approach

Another area of research associated with the production of intuitions is the "Dual Process" or "Dual System" account (Evans, 2010; Evans and Frankish, 2009; Evans and Over, 1996). According to this view, there are two types of processing architectures in the brain. On the one hand, there are judgments that involve explicit conscious reasoning and typically result in the formation of a belief. This is what happens when, say, someone deliberates on the premises of an argument and consciously forms a conclusion. This introspectively accessible process is said to be performed by "System 2" in the brain. This is contrasted with "System 1" psychological processes that are unconscious, more automatic, and typically outside of our control. Several authors have noted that it is likely that our intuitive responses to hypothetical cases in philosophy are generated by this latter, System 1 type of processing (Weinberg and Alexander, 2014; De Cruz, 2015). The processes that generate TEIJs are not introspectively accessible and appear to be quite automatic – we hear the scenario and just find that it strikes us a certain way. This has led at least some to propose investigating the psychological source of philosophical intuitions by using the dual-processing approach.

For example, DeCruz (2015) has employed psychologist Robert McCauley's account (2011) of two different sorts of Type 1 processing to explain a variety of philosophical intuitions, particular those involving teleology and epistemology. As De Cruz notes, McCauley distinguishes between what he refers to as "maturationally natural cognition" and "maturationally practiced cognition". The former involves Type 1 processing that spontaneously results from natural development, and requires no special training, such as learning to speak one's native language or ascribing mental states to others. The latter involves Type 1 processing that involves a distinct skill and requires some degree of training or practice, such as riding a bicycle or playing the guitar. De Cruz argues that both sorts of cognition likely give rise to philosophical intuitions. For instance, she suggests that intuitions involving teleology and knowledge possession arise from maturationally natural cognition (because these come about spontaneously without any special training) whereas intuitions about, say, how Hume would respond to a certain line of inquiry would be produced by maturationally practiced cognition (because it requires immersion into Hume scholarship). This is just one of the ways in which dual processing theories can be used to account for the production of philosophical intuitions.

C. The Heuristics Approach

A third area of research that has made a significant contribution to the etiological project is the important psychological work on various everyday heuristics that we use in a wide range of endeavors (Kahneman et al., 1982; Gigerenzer et al., 1999). Heuristics are often characterized as processing short-cuts or rules of thumb that are employed unconsciously to help us solve various problems. They tend to be faster and easier to use than more complicated calculations, but they can also be unreliable and can lead to errors. A classic example is the "recognition heuristic". Here, subjects employ the recognizability of some instance as an indicator of some other, less accessible property it might possess. For example, when asked which city is larger, it has been shown that people tend to use their recognition of the city's name as a guide to the size of the city – the more recognizable, the larger the city is assumed to be. This seems like a reasonable strategy since there is a rough correlation between the fame of a city and its size; however, it can lead subjects to make faulty inferences, even when they are provided with further information in particular cases suggesting that the recognizability of the city should be discounted (Gigerenzer et al., 1999).

One area of philosophical investigation that has explored the role of heuristics in producing philosophical intuitions is moral philosophy and psychology. Several authors have suggested that our moral intuitions are produced by moral heuristics that are employed in our moral judgments and often allow us to assign moral responsibility or determine a course of action. For example, Sinnott-Armstong et al. (2010) develop an account of moral heuristics and intuitions based upon Kahneman and Frederick's (2005) notion of unconscious attribution substitution. With unconscious attribution substitution, people unconsciously determine whether or not something has a certain attribute, F, by ascertaining whether or not it has some other attribute, G. People are typically unaware of substituting the heuristic attribute for the target attribute, and the substitution clearly involves some sort of tacit assumption about a correlation between the two. It is a useful heuristic because the target attribute is often difficult to detect, while the heuristic attribute is easier to reveal. Sinnott-Armstrong et al. go on to develop a sketch of how such a heuristic could give rise to moral intuitions. Because the moral rightness or wrongness of an act can often be very difficult to discern, they suggest that people substitute some other attribute, like what the majority is doing, or whether or not the act produces unpleasant feelings, to ascertain moral propriety. On this account, when presented with a thought experiment in ethics, the audience would intuitively judge what ought to be done by considering something else, like how the scenario makes one feel.

Summary

This is not intended to serve as an exhaustive overview of the only research programs that hold considerable promise of telling us about the source of TEIJs. To the contrary, it is intended to help illustrate just how much theoretical diversity exists with regard to the etiological project. And while these approaches are not always mutually exclusive, it is still true that, given the undeniable differences in these research agendas, this variance in theorizing is likely to contribute to the problem I will raise in the next section.

III. The Etiological-Diversity Problem

Before explaining the etiological-diversity problem, it will help to consider a couple of ways in which diversity regarding intuitions is *not* the worry I'm concerned about. One sort of diversity was mentioned in Section II – that a very broad range of different sorts of psychological states get referred to as 'intuition', including the seeming truthfulness of logical claims, semi-informed hunches, the puzzling nature of a paradox, and so on. But this diversity can be made far less problematic by simply stipulating that we restrict our analysis to one specific type of intuition, as I proposed in Section II. Thus, we can limit our focus to the sort of intuition that has played such an important role in theory development in philosophy – the immediate, spontaneous reaction to a philosophical thought experiment, what I've been calling TEIJs. Of course, TEIJs may still involve a range of different types of mental states, in which case my pessimism would only be reinforced. But for now, for the sake of this discussion, I'm going to assume a fair degree of uniformity among TEIJs, qua-mental state type. The second type of non-problematic diversity pertains to the fact that, as with nearly all faculties, the TEIJ-producing process probably involves a "boxology" of different sub-systems and modules that collectively generate the intuitive reaction. The mere existence of an array of contributing sub-modules poses no problem to developing a robust theory of how a given faculty works – indeed, spelling out the organizational mechanics or "flow-chart" of different cognitive process often provides one of the most satisfying explanations of how a cognitive task is performed.

If these forms of diversity do not pose a problem for the etiological project, then what sort of diversity does? Consider two different types of psychological states. For one type, there is something close to self-contained cognitive system or module that is responsible for creating the state in question. For example, our perceptual systems more or less work in a predictable and systematic manner to produce specific perceptual states such as visual or auditory experiences. We can study the auditory system and learn about its basic functioning – about the relevant parts of the brain that are involved and the sort of computational operations employed – and thereby gain a great deal of understanding about how hearing works. This understanding yields helpful predictions and generalizations because, although it is incredibly complicated, it is nevertheless the same core auditory system that is working throughout the wide array of conditions in which we use our hearing to navigate the world. While there are a variety of different stages and sub-systems involved, most of the time hearing involves the same sub-systems. Hence, there is a faculty of hearing that lends itself to comprehension, and thus allows for useful generalizations and predictions about the etiology of auditory experience.

However, not all psychological states have such an unvaried, self-contained faculty that is responsible for their production. Consider beliefs. Given the remarkably diverse array of cognitive processes and mechanisms that can generate beliefs – perception, memory, inference, introspection, etc. – it would clearly be a fundamental mistake to launch an inquiry into the "faculty of belief". That is, given the extreme diversity of cognitive processes that produce beliefs, we should not expect to find an account of belief production that will yield the sort of predictions and generalizations that we find with perceptual states. Learning about how beliefs are formed via a process of inference is not going to tell us very much about the way beliefs are formed via perceptual experience or recall. Of course, this doesn't mean there is no value to studying the cognitive etiology of beliefs, or that we can't gain useful information about these diverse processes of belief formation. But in the case of beliefs, it is clearly a mistake to ask questions such as whether or not *the* process that generates beliefs is reliable or improves with training or tends to malfunction in certain precise conditions. It is a mistake to ask such questions because there is no unique, uniform process that generates beliefs.

So, then, what about TEIJs? Traditionally, many have assumed that the proper model for understanding the etiology of intuitions is something like the perceptual model just presented. Many philosophers have treated rational intuition as if it was the a priori analogue of our senses, providing basic evidence about hidden philosophical truths, albeit in a non-sensory way (see, for example, Sosa, 1998; Hales, 2012). The intuition faculty is thereby treated as doing the same sort of work in philosophy that observation does in science – as providing a kind of access to extra-mental facts about ethics, epistemology, mind, language, and so on.

Naturalistically inclined philosophers, including those in experimental philosophy, have come to reject the idea that intuitions are produced by cognitive systems that provide a window or insight into the realm of abstract philosophical truths. They are instead committed to explaining the production of intuitions along the lines discussed in the last section – as a process of concept application or as involving the use of various heuristics. Yet, while there has been a rejection of the perception-like model, it is less clear that naturalistic philosophers have abandoned the related idea that there is a stable faculty or module that generates intuitions about particular philosophical matters. For example, Shieber (2012) suggests that there are domain-specific, hard-wired, functionally distinct modules that are similar to a linguistic processing module and that generate intuitions about the different domains of philosophy, such as ethics, epistemology and metaphysics. While Schieber's "Modularity Model of Intuition" stipulates that there are different modules corresponding to the different areas of philosophy, it implies that the cognitive structures delivering intuitions in each area are relatively stable, fixed and uniform.[4]

But it is far from clear that an even limited sort of uniformity thesis is correct. Instead, as I'll now argue, when we look closely at the way the etiological project is actually carried out, there is good reason to suppose that the array of operations and sub-mechanisms producing philosophical intuitive reactions, even in specific areas of philosophy, vary and diverge significantly. That is, the expectation that we will discover a stable module or set of modules that systematically generate TEIJs is overly optimistic. Given the way the research is unfolding, we should instead anticipate a mixed bag of varying procedures and mechanisms that produce TEIJs and that will preclude substantial generalizations and predictions.

[4] Similar perspectives have been suggested about epistemological intuitions by Nagel (2012) and moral intuitions by Hauser, Young and Cushman (2008).

As we saw in Section II, the range of different research programs that have contributed theories about the cognitive etiology of TEIJs is fairly broad. These different research programs, with their own theoretical posits, principles and approaches, create considerable theoretical diversity despite overlapping to some degree. McCauley's Type 1 account of cognition as a source of intuition involving 'maturationally natural' and 'maturationally practiced' cognitive processes is very different from, say, a prototype account of how the judgment is produced that appeals to a typicality ranking, or an account that invokes a specific heuristic. Of course, *theoretical* diversity does not entail *processing* diversity or *mechanism* diversity. Still, this diversity in the theoretical terrain at least suggests that we should not expect a single coherent etiological account any time soon. Moreover, there is reason to believe that these different theoretical frameworks are enjoying success because they are actually describing different mechanisms and processes that produce TEIJs. In other words, insofar as they disagree, these competing theoretical frameworks may nevertheless be correct because they are capturing multiple types of operations and processes that are actually devoted to generating different intuitions in different circumstances.

Suppose we limit our evaluation to just one of these research approaches, and focus solely on the concept application/categorization framework. Recall that with this approach, TEIJs are identified with categorization judgments, produced by the application of philosophical concepts. The processes that generate philosophical intuitions are the psychological processes associated with the employment and application of a concept. But even if we focus just upon the concept application perspective, there is good reason to expect a considerable amount of etiological diversity. As we noted in Section II, there are three popular models – prototype, exemplar, and theory models – about what concepts are, how they are stored and retrieved, and how they allow us to classify things. Of course, within each of these paradigms there are theoretical variations and sub-variants endorsed by different researchers. For instance, there are different versions of prototype accounts that involve diverse ways in which item features are invoked for categorization (Murphy, 2002). This is perfectly normal for thriving research programs where an array of alternative theories compete for consensus through predictive and explanatory success. Yet even if we ignore the sub-variants within in each paradigm, competition among the three accounts of concept application has not yielded a clear winner. Instead, the different theories seem to be enjoying considerable success because of different robust findings that uniquely support each particular model. For instance, the easy learnability and classification of items that possess more prominent and common features supports most feature-based prototype accounts. However, the finding that we sometimes classify more quickly atypical members that are nevertheless similar to those we have been exposed to through personal experience undermines the prototype view, but instead supports the exemplar theory.

Various other findings that suggest classification does not involve any sort of the comparison process lend credence to the theoretical account. And, of course, there are findings that serve as evidence for theories of concepts that are different yet from these three (Murphy, 2002, Machery, 2009).

This diversity of successful theorizing has led some, in particular Edouard Machery and Daniel Weiskopf, to endorse what is sometimes known as the "heterogeneity hypothesis" or "pluralism" about concepts (Machery, 2009; Weiskopf, 2009). The heterogeneity/plurality hypothesis is simply the view there is no single correct answer to what it is for something to be a concept. For most items (kinds, properties, relations, events, etc.) we actually possess different kinds of concepts, particularly those represented in the manner suggested by the three models presented above. In other words, on the heterogeneity/plurality hypothesis, there is psychological variance not only in terms of what the concept is about (e.g., dog vs. cat), but also in terms of the sort of knowledge structure that is involved. When I go to classify something as an instance of, say, a dog, in certain contexts I'll do so by employing a prototype representation of a dog, while in other contexts I'll do so by invoking an exemplar (stored memory of familiar dog) or some deeper theoretical knowledge about dogs. There is no one account that is right and that can yield straightforward generalizations about the process of categorization. In other words, on the heterogeneity/plurality hypothesis, the cognitive nature of concept application/categorization is varietal and (duh!) heterogeneous.

A robust defense of the heterogeneity/plurality thesis about concepts would take us too far afield, so I will invite readers to examine the arguments presented by Machery and Weiskopf and decide for themselves. Most of these arguments strike me as absolutely convincing and present the most compelling way of making sense of the divergent psychological findings regarding the nature of concepts.[5] And while Machery goes so far as to recommend abandoning talk of concepts altogether (since the heterogeneity hypothesis implies there is no conceptual natural kind), we do not need to embrace concept eliminativism for the heterogeneity/plurality hypothesis to have profound implications for the etiological project. If it is correct, the heterogeneity/plurality hypothesis would mean that insofar as philosophical intuitions prompted by thought experiments are psychologically generated by the application of philosophical concepts, their production would involve, as a group, a very diverse range of different cognitive processes.[6] It wouldn't

[5] I will only mention that at least some alternative accounts of the divergent findings, such as various types of hybrid accounts (for example, Osherson and Smith, 1981) partially agree with the heterogeneity/plurality thesis that different bodies of knowledge and processes can be involved in the application of a concept. For example, in Osherson and Smith's account, we might classify something by using a definition in some contexts and by using a prototype in other contexts. With such an account, there would also be diverse processes that generate TEIJs.

[6] Thus, my claim in this section is a conditional one: if the heterogeneity/plurality thesis is correct, it would mean that even if TEIJs are solely produced by processes of concept application, this would still involve a diverse range of very different processes generating TEIJs.

simply mean that the cognitive process that generates an intuitive reaction to a Gettier-type case might involve one type of concept, say a prototype concept of knowledge, while the process that results in an intuition about a Kripke-style Jonah case would involve a different type of concept, say a tacit theory about reference. It would also mean that the conceptual machinery employed in response to one sort of Gettier case could be substantially different than the conceptual machinery employed in a different sort of Gettier case. Indeed, since several different factors can trigger the employment of one type of concept rather than another, it could entail the retrieval of different types of concepts in different people with regard to the same Gettier case, or different types of concepts in the same person in response to the same case at different times in different settings. Learning that subjects A, B and C employed a certain sort of prototype in categorizing Gettier's 10 Coins Case as non-knowledge would be perfectly compatible with subjects D, E and F employing an exemplar concept in response to that case, or A, B and C employing an exemplar concept in response to a Fake Barn case. The amount of intuitional etiological diversity allowed by the heterogeneity/plurality hypothesis is such that it would undermine the hope of making broad generalizations about how, psychologically, humans respond to philosophical thought experiments. Even if we could somehow determine that, say, a particular process involving a prototype concept of knowledge yielded a reliable intuition about a Gettier case, we would not be able to generalize that assessment to Gettier intuitions more broadly given that it is unlikely that such a process is invoked universally. And remember, this is the range of etiological diversity that exists even after stipulating that our intuitive responses to philosophical cases are entirely due to the application of concepts. Considering that it is more likely that TEIJs are sometimes produced by processes other than concept application, such as the employment of a heuristic or some other form of Type 1 mechanism, then my pessimism about the etiological project providing useful generalizations looks even more warranted.

So far I've addressed the prospects of the etiological project from the perspective of some of the different research programs devoted to explaining the generation of philosophical intuitions. My claim has been that from the perspective of these different research agendas, it is highly doubtful that we will eventually end up with an all-encompassing, general account of intuition etiological processes and mechanisms or modules that can provide very helpful generalizations or that can answer the sort of broad questions many have about TEIJs. We are instead likely to find a diverse array of idiosyncratic and highly specific etiological processes that apply only to particular intuitive judgments in particular situations. This verdict is substantially reinforced when, instead of looking at research paradigms broadly construed, we look instead at the way investigators try to account for specific intuitive judgments that are reactions to specific thought experiments. When we look at proposals about how individual sorts of TEIJs are produced, we find very particular, idiosyncratic accounts positing operations that make sense only as theories about the specific sort of intuitive judgment in question. Theories about the way, say, certain intuitions about moral propriety are produced posit such detailed and specific machinery that there is very little chance that the same sort of machinery has much to do with the way intuitions about epistemology or metaphysics are produced. Indeed, even within particular areas of philosophical specialization, like ethics or epistemology, the cognitive processes invoked to explain specific philosophical intuitions prompted by one sort of case often have little in common with the cognitive processes invoked to explain other philosophical intuitions about other sorts of cases in the same area. In a sense, this shouldn't be terribly surprising. Researchers construct models that are fine-tuned to explain idiosyncrasies and specific noteworthy findings associated with individual TEIJs. In the process, they typically propose functional elements and operations that would generate the particular observed effects associated with the intuitive reaction, and that make sense of only certain distinctive intuitional phenomena.

To see this better, consider one of the most active areas of intuition etiological research – work regarding the attribution of intentional action and on what has come to be called the "side-effect effect". Joshua Knobe (Knobe, 2004, 2008) has shown that people's intuitions about whether or not an act was performed intentionally is often strongly influenced by their moral evaluation of the act. Subjects are presented with two scenarios: one where the chairman of company adopts a policy for economic reasons but that he knows will also harm the environment, and one, (worded in nearly identical language) where he knows it will have the side-effect of helping the environment. Remarkably, 82% of the subjects intuited that the vice president intentionally harmed the environment, but only 23% said he intentionally helped the environment. It seems that the subjects' moral assessment of an act (harming as opposed to helping) strongly influenced their views on the actor's intentions.

Because this is a surprising finding about an intuitive response, there has been considerable speculation about the underlying psychological process that creates this TEIJ. For our purposes, what matters is the specific nature of cognitive models that have been proposed. For example, Knobe has argued that the reaction in the harm case is due to a subtle interplay of perceived features of the case that (unconsciously) play a critical role in the formation of the intuition of intentionality. The process is one where the subject first determines if the behavior is bad or good, and then assigns intentionality depending on the presence of differing features, such as the result being foreseeable by the agent or involving some degree of effort, depending on the specifics of the case. On this etiological model, "people's intentional action intuitions tend to track the psychological features that are most relevant to praise or blame. But – and this where the moral considerations come in – different psychological features will be relevant depending on whether the behavior itself is good or bad" (Knobe, 2008, p. 144).

The critical thing to note here is how specific this account is to the particular sort of TEIJ that is being explained. This account makes sense only with regard to intuitions about whether or not an act is performed intentionally by someone in a specific vignette. The account has little value in explaining the production of other philosophical intuitions, such as intuitions about knowledge, personal identity, reference, or even intuitions about deliberate action that are prompted by scenarios significantly different from the harm/help scenarios. That's because the model is designed to account for a very particular intuition, and is in no way intended as some sort of general purpose model. Indeed, even within Knobe's account, there is considerable variance. As Knobe puts it, "...a given feature may be highly relevant to the praise or blame the agent receives for one behavior while remaining almost entirely irrelevant to the praise or blame the agent receives for another, somewhat different behavior...People's intentional action intuitions seem to exhibit certain flexibility, such that they look for different features when confronted with different behaviors..." (2008, 143–144). In other words, the specific process that generates an intentionality intuition is going to vary from case to case.

When we look at the models designed to explain other philosophical intuitions, we find the same sort of idiosyncratic functionality (see for example, Arico, et. al., 2011; Sytsma and Machery, 2009). Between these accounts, we see very little overlap, except in the boring sense in which they all attempt to explain an intuitive response by appealing to mechanisms and processes that "make sense" of that particular intuition. Beyond that, these accounts have about as much in common as psychological accounts of inference and short-term memory – both processes that generate beliefs. For most experimental philosophers engaging in the cognitive program, explanatory breadth and broad-ranging principles of the sort that would allow noteworthy generalizations are neither essential nor expected. Given this trend in explaining where TEIJs come from, it seems quite unlikely that we will get any sort of unified theory about a faculty that is *the* source of philosophical intuitions.

To wrap up this section, consider again one of the central sort of questions it is hoped the etiological project will answer: "Are the cognitive processes and mechanisms that produce philosophical intuitions in response to a thought experiment adequately reliable?" It is becoming increasingly clear that the answer no doubt depends upon *which* philosophical intuitions and *which* cognitive process and mechanisms you are talking about. Indeed, the answer may depend upon which cognitive processes and mechanisms producing which intuitions, in whom, in what sort of setting. There really is no reason to think that we will eventually develop a uniform account of something like an intuition faculty that can answer broad questions about the general nature of TEIJs.

IV. The Secondary Calibration Problem

In the last section, I raised doubts about our ability to find general answers to important questions about the etiology of our intuitive reactions to philosophical thought experiments – doubts based upon signs pointing to the extreme diversity of processes and mechanisms that can give rise to TEIJs. In this section, I'm going to raise another problem that would exist even without this high degree of etiological diversity. As we saw in Section II, the questions we seek to answer about intuitions by investigating their etiology have a normative dimension to them. Are the intuitions produced trustworthy? In what sort of contexts are they likely to go awry in the production of intuitions? The etiological project is at least partially based upon the hope that a deeper understanding of the cognitive processes and mechanisms that generate intuitions will help us to answer these questions.

This hope is largely motivated by the recognition of a difficulty in evaluating the truthfulness of philosophical intuitions. As a number of authors have pointed out, perhaps most notably Robert Cummins (1998), we do not have an independent method for assessing the truthfulness of the contents of these intuitive reactions; we only have the intuitive reactions themselves. It may seem true that in Gettier's 10 coins case Smith's belief is not knowledge, but, ironically, it is far from clear how we can know if this intuitive reaction is correct. As Cummins notes, processes and mechanisms that deliver information need to be properly calibrated, and proper calibration requires independent access to the sort of facts or state of affairs that the information is about. But in most cases, there is no independent access to the philosophical matters that our intuitive reactions are about. This is somewhat unique to TEIJs. If something perceptually seems a certain way to us, like the lines appearing to be of different length in the Mueller-Lyer illusion, there are strategies we can employ to establish that this seeming is mistaken, that the lines are actually the same length (we can, for example, measure them). If we want to test the reliability of memories in certain contexts, we can do things (like, say, use audio-video recordings) to provide an independent mode of access to the relevant source of information, and then compare this to the memories. But what do we examine to test the sort of metaphysical, epistemological and ethical intuitions that are generated by the presentation of a thought-experiment? What can we do to test whether the common reaction to a Gettier case reveals a deep fact about knowledge, as opposed to a widespread delusion about knowledge? With regard to the way things intuitively seem to us in response to many philosophical thought experiments, there is no real strategy for assessing those intellectual seemings beyond the philosophical theories that the intuitions are employed to support or challenge. Reflective equilibrium is often presented as a process whereby we *can* do something like calibration – weighing our commitment to the intuition against theoretical commitments with which they conflict.

Yet, as Cummins and others have noted, we very rarely reject an intuition because it conflicts with a philosophical theory. Instead, we almost always use the intuition as compelling evidence that a given theoretical stance is false. This has come to be known as the "calibration problem" with regard to philosophical intuitions.

So one of the reasons we would like to learn more about the mechanisms and processes that generate TEIJs is the hope that it might help provide a solution to the calibration problem. If we can come to understand how the intuitions are generated, then we might gain a better understanding of their reliability. As De Cruz puts it, "...the psychological underpinning of intuitions can provide a test key with which to test their validity" (2015, p. 236). Similarly, Knobe and Nichols propose that, "[f]irst we use the experimental results to develop a theory about the underlying psychological processes that generate people's intuitions; then we use our theory about the psychological processes to determine whether or not those intuitions are warranted" (2008, p. 8).

However, there is a fundamental problem with this approach. The standard way we normatively evaluate the functioning of a cognitive process or mechanism is to first evaluate the quality of the cognitive states it produces in different modes and circumstances; then, we use that assessment to evaluate the quality of the producer in those different modes and circumstances. In other words, we evaluate the normative status of the cognitive mechanisms and processes by virtue of what they produce. For instance, we determine how well a sensory system is functioning in different circumstances by examining the quality of the perceptual states it produces under those conditions. It is functioning properly when it produces normatively appropriate (e.g., accurate, sufficiently clear, timely, etc.) perceptual states. And it can often be characterized as malfunctioning (or irrational, or a flawed heuristic, etc.) when it produces states that are not normatively appropriate – states that are false or irrational or in some other way inadequate. We gain a sense of how reliable a cognitive system is in various contexts by first ascertaining the proportion of truthful or accurate states it produces in that context. And this, of course, requires us to be able to first assess whether or not a state that is produced is actually true or false.

Consequently, because the evaluation of the truthfulness or normative value of TEIJs is deeply problematic, the normative status of the cognitive processes and mechanisms that produce those intuitive judgments is also deeply problematic. Cummins' calibration problem is generally regarded as a problem with calibrating the intuitions themselves – with being able to assess the truthfulness of their contents. However, we can now see that it extends beyond our evaluation of the TEIJs. It also applies to the evaluation of the processes that produce those intuitions because we have no clear way to calibrate the functioning of these processes without first calibrating their output. In other words, what I am calling the secondary calibration problem is based upon the primary calibration problem – the fact that we lack independent access to philosophical truths prevents us from evaluating the accuracy of philosophical intuitions. Because we cannot evaluate the accuracy of the intuitions, we can't normatively evaluate the cognitive mechanisms that generate them. So the hope of many, that we can overcome the calibration problem by learning more about the etiology of TEIJs, is deeply misguided because it ignores the fact that the normative assessment of cognitive producers usually requires a prior assessment of the cognitive product – the very thing that is missing in the case of TEIJs.

The findings of experimental philosophy exacerbate the secondary calibration problem because they reveal considerable variance in the intuitive reactions to particular vignettes and thereby raise the question of how we can know which of these different intuitive reactions are correct. But even if this variance had not been demonstrated, or even if widespread agreement had been shown, we would still have a legitimate concern that this agreement reflects nothing more than a systematic delusion. We already know that in the case of several cognitive processes, including different problem-solving strategies, people predictably and systematically generate faulty and inaccurate mental states (see, for example, Nisbett and Ross, 1980; Kahneman, Slovic and Tversky, 1982). We know that many of our cognitive faculties systematically fail to produce normatively correct intuitions, especially in particular contexts. But we know this only because we can test the mental states produced in those contexts. We can measure the lines in a Mueller-Lyer drawing and show that our visual seemings are incorrect. We can do careful statistical analysis to show that widespread intuitions about probabilities are mistaken. Using these findings, we can *then* theorize about the normative nature of the cognitive machinery that produces these intuitions. But without these findings about their output, we would have no basis for evaluating their functioning. And it is precisely this sort of finding about the status of TEIJs that we don't have. Consequently, we have no basis for evaluating the functioning of the cognitive mechanisms that generate TEIJS, and thus we cannot use such an evaluation of the functioning of those cognitive mechanisms to provide an assessment of those TEIJs.

Another way to think about this issue is in terms of Chomsky's well-known distinction between competence and performance (Chomsky, 1965). Chomsky and other linguists argue that an underlying linguistic competence involving some sort of knowledge structure, like a grammar, produces much of our linguistic output, including linguistic intuitions. However, in certain situations and contexts, our actual linguistic performance does not properly reflect the nature of this underlying competence because of performance errors, caused by influences or limitations of other important sub-systems, such as working memory or attention. Note that we can call the output in such situations performance errors because we have a method for evaluating the status of the linguistic output. We can, for example, describe the faulty intuition that sentence (A) is ungrammatical by carefully examining it and seeing that, in fact, it properly follows the rules for multiply-center-embedded clauses, albeit in a very complicated way:

(A) A man that a woman that a child that a bird saw knows loves.

We can then speculate about the cognitive source of this intuitional error, and assign it to something like a failing in working memory. In any event, this breakdown in the production of linguistic intuition can only be ascertained once we have first determined the mistaken nature of the intuition itself. And this is only possible because we have an independent source of information (English grammar) about center-embedded clauses that reveals the flawed nature of the intuition. If we had no way to assess the accuracy of linguistic intuition, and all we had to go on was the intuition itself, then it would be very difficult to see how we could distinguish between processes and systems that contribute to linguistic competence, or the expression of linguistic competence, and processes that instead contributed to some sort of intuitional error. Any talk of systematic performance errors presupposes the ability to determine flawed output. And that ability appears to be exactly what is missing in the case of TEIJs.

One possible strategy for addressing this issue and solving the calibration problem has been proposed by Talbot (2009) and is at least suggested by De Cruz (2015).[7] The proposal is that we can learn about the reliability of processes generating intuitions on mundane matters for which we actually *do* have independent access to the truth. We can then extend what we learn to the more enigmatic TEIJs and how they are produced. As Talbot puts it:

> We can determine how intuitions work – the data they are likely to be sensitive to and the data they are likely to ignore, and what factors make them more or less accurate – by studying intuitions about non-philosophical questions we know the correct answers to. These are questions about ordinary objects, behavior, possibilities and so forth. We can compare what we learn about how intuitions do and do not work for ordinary questions with our demands on a source of evidence for philosophical questions, and calibrate our intuitions in this way (2009, p. 165).

Similarly, De Cruz suggests that if Gettier intuitions arise from folk psychological mechanisms, as suggested by Nagel (2012), then we can evaluate the accuracy of those intuitions by generalizing from the accuracy of folk psychological intuitions on more mundane, testable matters (De Cruz, 2015, p. 9).

While I regard these proposals as interesting, there are two reason why I am not very optimistic about using this strategy for evaluating philosophical intuitions and their psychological origins. The first is that a well-known source

[7] I am grateful to an anonymous reviewer for correctly pointing out that I should address Talbot's and De Cruz's proposals.

of cognitive error is the employment of cognitive processes or mechanisms in contexts that are different from those in which they function properly. As Nisbett and Ross put it in their seminal text on cognitive errors: "We contend that people's inferential strategies are well adapted to deal with a wide range of problems, but that these same strategies become a liability when they are applied beyond that range..." (Nisbett and Ross, 1980, xii). This point applies to a variety of cognitive mechanisms and operations, including whatever psychological systems and sub-systems that generate intuitions. Just because a cognitive strategy produces reliable intuitions about some ordinary testable matter (such as, say, whether or not a physical object could be used as a hammer), this gives us no reason to think the same strategy would be reliable – or even work at all – to create true intuitions about the nature of knowledge, free action, reference or any of the other abstract matters philosophers care about. Indeed, as Machery has recently argued (Machery 2017), philosophical cases are more prone to generating intuitional errors because they typically involve judgments about bizarre scenarios with "disturbing characteristics", like highly unusual circumstances and separating properties that normally go together. Intuitions about the ordinary can't go very far in helping us calibrate philosophical intuitions precisely because the latter are almost always about the extraordinary.

The second problem with this proposal stems from the etiological diversity challenge we discussed in the last section. Talbot and De Cruz's calibration method assumes a high degree of uniformity in the production of intuitions, since the processes and mechanisms need to be relatively constant and invariable for the lessons learned about intuitions in one domain to apply to others. But as we saw in Section III, this sort of uniformity assumption is highly dubious.[8] Not only is it doubtful that the exact same cognitive machinery that generates intuitions about ordinary matters also serves to generate intuitions about philosophical cases, but it is also doubtful that the cognitive machinery that generates intuitions about particular philosophical cases in certain contexts is the same machinery that generates intuitions about other philosophical cases or even similar philosophical cases in other contexts. This proposed solution to the calibration problem requires precisely the sort of generalizations about the etiology of intuitions that the etiological diversity problem prevents.

It might be supposed that we can evaluate the reliability of a cognitive process that generates a philosophical intuition by demonstrating interference – causal influence by sub-systems that have been independently determined to be inappropriate. Through independent investigations, we might be able determine that a certain module functions to generate a particular sort of psychological output, such as an emotional response or a certain

[8] I'm grateful to an anonymous reviewer for pointing out the relation between the etiological diversity problem and Talbot and De Cruz's arguments about calibrating intuitions.

attitude. Then perhaps it could be argued that this output is inappropriately influencing the TEIJ in question, and thus the process could be characterized as introducing a corrupting, error-producing factor. Yet the problem with this strategy is that it is always possible for the defenders of the intuition to argue that the module's output is not inappropriate at all, and that, in fact, the error lies in any theory that claims otherwise. Indeed, a philosophical theorist that embraces the intuition can say these findings reveal an important deep truth about the philosophical issue in question; namely, that contrary to what some might have thought, it turns out that truthful intuitions about subject X require precisely the allegedly "corrupting" sort of input. It seems we can't get away from the fact that any treatment of a cognitive process as corrupting requires the prior assumption that the output it helps generate is, in fact, in error. One theorist's evidence of performance error is another theorist's evidence of underlying competence.

A nice illustration of this point involves a well-known argument by Joshua Greene challenging deontological ethical theories as an attempt to rationalize flawed moral intuitions (Greene, 2008). Greene argues that for vignettes describing certain forms of up-close and personal violence, regions of the brain known for emotional reactions become highly activated. The amygdala, for instance, which is known to be involved in the production of strong emotions, is highly activated during footbridge versions of the Trolley case (when most subjects have the intuition it is morally wrong to personally cause a death to save many lives – e.g., by pushing a man off a bridge), but it is mostly inert when the subject is presented with switch versions of the case (where most subjects have the intuition it is morally permissible to do something remote – throw a switch – that will cause the one death to save many lives). This leads Greene to suggest that the amygdala is introducing a corrupting influence of emotion in the footbridge intuitions. Greene argues that the proper moral intuition, reflecting genuine moral reasoning, is the utilitarian response exhibited in the switch case. Although deontological theories (at least ones that argue it is always wrong to cause the death of an innocent) are supported by the footbridge intuitions, Greene argues these intuitions are not proper moral reactions because of the corrupting influence of emotion.

In response to Greene's indictment of intuitions supporting deontology and opposing utilitarianism, some authors have directly challenged Greene's claim that emotional input from brain regions like the amygdala should be regarded as corrupting to moral judgment.[9] Instead, people within the "sentimentalist" tradition of deontology (or, indeed, any form of sentimentalist moral theory) argue that the production of proper moral intuitions *requires* an emotional component. For them, the influence of the amygdala is exactly what one should expect to find when people engage in proper moral reasoning and reflection. Without any independent and compelling means by which we

[9] See, for example, Timmons (2008).

could evaluate and characterize the footbridge intuitions as flawed, Greene has no real basis for treating an emotion-generating cognitive mechanism like the amygdala as having an undesirable influence on intuition.

Consider again Knobe's interesting account of the mechanisms generating intuitions about intentional action. An important feature of Knobe's model is that the introduction of moral judgment in the process is no longer treated as a corrupting influence, but instead as part of our underlying competence in assigning intentionality. As Knobe puts it, "[t]he chief contribution of this new model is the distinctive status it accords to moral considerations. Gone is the idea that moral considerations are 'distorting' or 'biasing' a process whose real purpose lies elsewhere. Instead, the claim is that moral considerations are playing a helpful role in people's underlying competence itself" (2008, p. 145). Here again we see how a process that some might characterize as corrupting can be turned around as a proper part of our competence. This will always be possible because without an independent assessment of the status of the intuitions produced, we have no independent arbiter of what counts as corrupting vs. proper in the etiological process.[10]

This is not to say that we can't ever properly treat an *input* to the cognitive process as corrupting or contaminating to the intuition. In fact, experimental philosophy has revealed all sorts of ways in which we can show that TEIJs can be heavily influenced by factors that, normatively, ought not to have such an influence. For example, as Nichols and Knobe (2008) note, studies by Lerner et al. (1998) reveal that subjects who have their anger aroused by reading a maddening story about a bully are more inclined to assign responsibility to agents in unrelated cases of negligence. We can plausibly say that one *ought not* to judge others as more responsible for an act just because one is in an angry state, triggered by contemplation of an unrelated scenario. There is nothing problematic about identifying so called

[10] Nichols and Knobe (2008) consider this sort of no-arbiter problem in discussing intuitions about free will. As they note, studies reveal that subjects are more inclined to have compatibilist intuitions toward a vignette that is likely to produce strong affect (like someone murdering his family) and incompatibilist intuitions for low affect cases (like someone cheating on his taxes). The question is, which of these intuitions are "legitimate": in their terminology, there are two possible models – the "performance error model" and the "affective competence model", and it is far from clear which one is correct. They offer a clever study involving deterministic and non-deterministic universes to argue that the legitimate intuitions are the incompatibilist ones and, thus, the input of strong emotion (creating compatibilist intuitions) is corrupting. While their argument is compelling, I see no reason why an equally clever defender of compatibilism couldn't counter that affect is essential for proper intuitive judgments about free will, and that without it other corrupting influences (such as, say, tacit beliefs about agent causation or some such) bias the process when there is consideration of a deterministic universe. A similar concern is raised by Alexander, Mallon and Weinberg (2014), who point out that, "a proponent of an affective competence model could suggest that people's answers in the high-affect cases are fine, but that some other mechanism interferes with people's judgments in the low-affect cases..." (p. 41).

framing effects – ways in which it can be experimentally demonstrated that an intuitive reaction is influenced by factors that ought not to make a significant difference, like having watched a funny skit from Saturday Night Live (Valdesolo and DeSteno, 2006), or having the paper presenting the vignette sprayed with Lysol (Tobia et al., 2013). We can, in other words, employ a prior understanding of what should and should not be relevant to the determination of certain philosophical matters to treat specific inputs as contaminating if they influence a judgment about those matters.

It is important to note, however, that many of these revelations of contaminating or corrupting factors are not based upon the normative assessment of the cognitive machinery that produces the TEIJs. In accounts such as Lerner et al.'s, there is relatively little known about the actual cognitive machinery generating the judgment. Instead, these evaluations are based upon prior beliefs about the sort of factors that are relevant to the type of TEIJ generated. We say an intuition about the responsibility of a suspect should not be influenced by being in a state of anger caused by something completely unrelated to the crime because we have beliefs and attitudes about guilt that make such an influence illegitimate. The identification of framing effects is not an assessment of the output (TEIJ) based upon an assessment of the machinery that produces it. It is an assessment of the output based upon an assessment of the *input* to the machinery that produced it – that such input ought to be treated as irrelevant. Further understanding of the cognitive etiology is unlikely to have much bearing on that assessment.

Couldn't we use the assessment of corrupting input to impugn certain mechanisms or sub-systems, and then use the activation or engagement of such mechanisms to impugn the TEIJ? We might, for instance, show that a certain brain region becomes highly activated whenever an unusually disgusting scenario is presented, and then argue that it is something like a 'disgust affect generator'. Then, if such a mechanism is shown to contribute to a TEIJ in which a feeling of disgust should be irrelevant, we could use the influence of such a mechanism to challenge the epistemic status of the TEIJ. I don't want to claim that such a scenario is impossible, so I'll allow that this is perhaps a way in which the etiological project could show that a TEIJ is faulty. However, I suspect that this would really be another case where the inappropriate input would be providing the evidence of corruption, not an evaluation of the mechanism. For suppose that the mechanism was activated without the relevant input. One might then think that this would be an incontrovertible case where we could legitimately say a cognitive sub-system is corrupting the intuition-generation process (and not the input itself). However, it strikes me as more likely that one of two things would occur: Friends of the intuition would either argue that since the mechanism is active without being prompted by disgusting input, the mechanism actually *isn't* simply 'disgust affect generator' after all. *Or* the case reveals that, surprisingly enough, a little disgust is needed for the appropriate intuitive response! Either way, the case would be far from incontrovertible.

To sum up this section, Cummins' original calibration claim is about the psychological states themselves, the intuitings (the TEIJs). His claim is that we have no way to calibrate them, no independent way to assess the truthfulness of their contents. In reply, it has been suggested that perhaps we *can* calibrate them by closely examining the cognitive mechanisms that produce them and then use our evaluation of the mechanism's reliability to assess the TEIJs. And my rebuttal to that claim is that this won't work because, normally, we first need an assessment of the *output* of cognitive mechanisms to determine the mechanisms' reliability; we can't, in other words, put the mechanisms-assessment cart before the output-assessment horse. Cummins' calibration problem cannot be resolved by looking at the cognitive mechanisms and processes that produce TEIJs because the calibration problem extends to our evaluation of those cognitive processes and mechanisms themselves. Normative assessments of cognitive processes typically require normative assessment of their outputs, and that is precisely what is missing given the original calibration problem. While perhaps there will be some limited ways in which a cognitive process can be impugned as corrupting, I suspect these will be rare, given that friends of the intuition can always claim that the allegedly corrupting element is really part of our underlying competence. Finally, it should be noted that the most of these strategies are ones that could be used to show that an intuition-producing process is flawed or in error. I have no idea how we could demonstrate that an etiological process is reliable and produces appropriate intuitions. Given that one of the main goals of the etiological program is to show if and when intuitions are accurate, this is a major limitation.

V. Conclusion

Assuming the arguments presented above are sound, what should be the way forward with regard to our investigation of the etiology of philosophical intuitions? While I have offered a somewhat pessimistic outlook about what such an investigation is likely to yield, I certainly do not want to imply that I think investigating the psychological source of philosophical intuitions should stop or is a worthless enterprise. Quite the contrary – I believe that there is a great deal to be learned that will be valuable, and that although we may be unable to answer all of the original questions we began with, we can still address a number of important ones.

In light of the etiological diversity problem, how should we proceed? Although I believe we need to substantially temper or even deflate our expectations for the etiological project, the good news is that we can and should continue doing much of what we saw in Section III is already being done by many researchers; namely, rather than looking for a psychological account of *the* faculty of intuition, instead investigating how particular individual TEIJs are created in particular situations and with particular populations. Rather than trying to find an account that explains how epistemological intuitions or metaphysical intuitions are generated, we should instead focus on explaining individual reactions to specific thought experiments, trying to construct a plausible model tailored for each one. Perhaps in certain cases the application of some sort of heuristic makes the most sense, whereas in others it might involve the application of a specific type of concept, whereas with others maybe some other sort of complex process is involved. This means there won't be *an* etiological project, but many. While investigating the cognitive source of each individual sort of TEIJ may sound daunting, the list of important and impactful intuition-producing thought experiments in philosophy is not terribly long. And as we saw, much of this sort of work is already underway in the cognitive program of experimental philosophy, especially in the area of moral psychology.

What about the secondary calibration problem? As I noted in section IV, there may be some clever ways in which the reliability of a TEIJ can be called into doubt by focusing on cognitive machinery producing it. But I am inclined to think this will be the exception and not the norm. And while the unreliability of intuitions can occasionally be shown, I am much less optimistic about the prospects of demonstrating the trustworthiness of intuitions in this way. Thus, this problem adds to a growing list of considerations that have motivated several authors to challenge the import we typically assign to philosophical intuitions prompted by thought experiments.[11] Given these considerations, it makes sense to step back and ask what we want an account of some philosophical notion to do for us. Traditionally, we want that account to accurately capture some sort of mind-independent deep philosophical truth, and the hope was that intuition would guide us to such an account. However, if there is no way to demonstrate how and when the cognitive machinery producing TEIJs succeeds in relaying or transmitting these philosophical truths, then perhaps it is time to reconsider how we assess our philosophical theories.

One sensible option is to recognize that philosophical notions like knowledge, justice, freedom, responsibility, reference, moral goodness and the like all play very important roles in different aspects of our lives, including social institutions, legal systems, politics, science, normativity and so on. Instead of treating TEIJs as something like observational evidence, and thereby treating them as of central importance in determining which philosophical accounts we ought to embrace, we can and probably should diminish their significance and allow other, more pragmatic considerations take center stage. We should favor theories not because they are sanctioned by intuition, but because they put forward philosophical principles and concepts that do the sort of work we need them to do. Take, for instance,

[11] See, for example, Machery (2017), Weatherson (2003) and Weinberg, Nichols and Stich (2001).

our conception of knowledge and the Gettier intuitions that challenge the justified, true belief account. Given that the justified true belief account is straightforward, easy to use, and works quite well in distinguishing knowledge from non-knowledge in nearly every important and realistic setting (such as in distinguishing science from non-science), and given that we have no way of ascertaining the accuracy of the Gettier intuitions, a reasonable response would be to simply ignore them (Weatherson, 2003). This is just one example of how, although the etiological project cannot inform us about a faculty of intuition (because there is none), nor conclusively show that intuitions are produced in a reliable manner, it can nevertheless motivate a sensible path forward by diminishing our reliance upon intuitions. Indeed, the problems that plague the etiological project may eventually prove to be benefits insofar as they help promote a more reasonable and pragmatic outlook on theory development in philosophy.

References

Alexander, J., 2012, *Experimental Philosophy: An Introduction*, Malden, MA, Polity Press.

Alexander, J., Mallon, R. and Weinberg, J., 2014, "Accentuate the Negative", in J. Knobe and S. Nichols (eds.), *Experimental Philosophy, Volume 2*, Oxford, Oxford University Press.

Alexander, J. and Weinberg, J., 2014, "The 'Unreliability' of Epistemic Intuitions", in E. Machery and E. O'Neill (eds.), *Current Controversies in Experimental Philosophy*, New York, Taylor and Francis, pp. 128–145.

Allman, J. and Woodward, J., 2008, "What are Moral Intuitions and why Should We Care About Them? A Neurobiological Perspective", *Philosophical Issues*, 18, pp. 164–185.

Arico, A., Fiala, B., Goldberg, R., and Nichols, S., 2011, "The Folk Psychology of Consciousness", *Mind and Language*, Vol. 26:3, pp. 327–352.

Barsalou, L. W., Huttenlocher, J., & Lamberts, K., 1998, "Basing Categorization on Individuals and Events. *Cognitive Psychology*, 36, pp. 203–272.

Bealer, G.,1998, "Intuition and the Autonomy of Philosophy," in M. DePaul and W. Ramsey, *Rethinking Intuition*, Lanham, MD, Roman and Littlefield, pp. 201–240.

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

Cummins, R., 1998, "Reflections on Reflective Equilibrium," in M. DePaul and W. Ramsey, *Rethinking Intuition*, Lanham, MD, Roman and Littlefield, pp. 113–128.

De Cruz, H., 2015, "Where Philosophical Intuitions Come From", *Australasian Journal of Philosophy*, Vol. 93, No. 2, pp. 233–249.

Earlenbaugh, J. and B. Molyneux, 2009, "Intuitions are Inclinations to Believe," *Philosophical Studies*, 145, pp. 89–109.

Evans, J. S. B. T., 2010, *Thinking Twice: Two Minds in One Brain*, Oxford: Oxford University Press.

Evans, J. S. B. T. and Frankish, K., 2009, *In Two Minds: Dual Processes and Beyond*, Oxford: Oxford University Press.

Evans, J. S. B. T. and Over, D., 1996, *Rationality and Reasoning,* Hove, UK: Psychology Press.

Fischer, E. and Collins, J., 2015, "Rationalism and Naturalism in the Age of Experimental Philosophy", in E. Fischer and J. Collins (eds.), *Experimental Philosophy, Rationalism and Naturalism: Rethinking Philosophical Method*, London, Routledge Press, pp. 3–33.

Gigerenzer, G., Todd, P. M., & the ABC Research Group, 1999, *Simple Heuristics that Make us Smart*, New York: Oxford University Press.

Goldman, A., 2007, "Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status," *Grazer Philosophische Studien*, 74, pp. 1–26.

Greene, J., 2008, "The Secret Joke of Kant's Soul", in W. Sinnott-Armstrong (ed.) *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, Cambridge, MA, MIT Press, pp. 35–80.

Hales, S. D., 2012, "The Faculty of Intuition", *Analytic Philosophy*, Vol. 53:2, pp. 180–207.

Hauser, M., Young, L., and Cushman, F., 2008, "Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions", in W. Sinnott-Armstrong (ed.), *Moral Psychology, Volume 2 – The Cognitive Science of Morality: Intuition and Diversity*, Cambridge, MA, MIT Press, pp. 107–143.

Kahneman, D. and Frederick, S., 2005, "A Model of Heuristic Judgment", in K. Holyoak and R. Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning*, New York: Cambridge University Press, pp. 267–293.

Kahneman, D., Slovic, P. and Tversky, A. (eds.) 1982, *Judgment Under Uncertainty: Heuristics and Biases,* Camrbidge, Cambridge University Press.

Keil, F. C., 1989, *Concepts, Kinds, and Cognitive Development*, Cambridge: MIT Press.

Knobe, J., 2004, "Intention, Intentional Action and Moral Considerations", *Analysis*, 64, pp. 190–193.

Knobe, J., 2008, "The Concept of Intentional Action", in J. Knobe and S. Nichols (ed.) *Experimental Philosophy*, Oxford, Oxford University Press, pp. 129–147.

Knobe, J. and Nichols, S., 2008, "An Experimental Philosophy Manifesto", in J. Knobe and S. Nichols (ed.) *Experimental Philosophy*, Oxford, Oxford University Press, pp. 3–16.

Knobe, J. and Prinze, J., 2008, "Intuitions About Consciousness: Experimental Studies", *Phenomenology and Cognitive Sciences* 7, pp. 67–85.

Lerner, J., Goldberg, J., and Tetlock, P., 1998, "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility", *Personality and Social Psychology*, Bulletin 24.

Lewis, D., 1983, *Philosophical Papers: Volume I*, New York: Oxford University Press, pp. x.

Lycan, W., 1988, *Judgment and Justification*. Cambridge: Cambridge University Press.

Machery, E., 2009, *Doing Without Concepts*, New York, Oxford University Press.

Machery, E., 2017, *Philosophy Within its Proper Bounds*, Oxford, Oxford University Press.

McCauley, R., 2011, *Why Religion is Natural and Science is Not*, Oxford, Oxford University Press.

Murphy, G., 2002, *The Big Book of Concepts*, Cambridge, MA, MIT Press.

Nado, J., 2013, "Why Intuition"? *Philosophy and Phenomenological Research,* 86, pp. 15–41.

Nagel, J., 2012, "Intuitions and Experiments: A Defense of the Case Method in Epistemology", *Philosophy and Phenomenological Research*, Vol. 85: 3, pp. 495–527.

Nichols, S. and Knobe, J., 2008, "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions", in J. Knobe and S. Nichols (ed.) *Experimental Philosophy*, Oxford, Oxford University Press, pp. 105–126.

Pust, J., 2017, "Intuition", *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/intuition/>.

Rosch, E., 1975, "Cognitive Representation of Semantic Categories", *Journal of Experimental Psychology*, v. 104, pp. 192–233.

Scholl, B. 2007, "Object Persistence in Philosophy and Psychology", *Mind and Language* 22, pp. 563–591.

Shieber, J., 2012, "A Partial Defense of Intuition on Naturalist Grounds," *Synthese*, 187: 2, pp. 321–341.

Sinnnot-Armstrong, W., Young, L. and Cushman, F., 2010, "Moral Intuitions" in J. Doris (ed.) *The Moral Psychology Handbook*, Oxford, Oxford University Press, pp. 246–272.

Smith, E. and Medin, D., 1981, *Categories and Concepts*, Boston, Harvard University Press.

Sosa, E., 1998, "Minimal Intuition," in M. DePaul and W. Ramsey, *Rethinking Intuition*, Lanham, MD, Roman and Littlefield, pp. 257–269.

Sytsma, J. and Livengood, J., 2016, *The Theory and Practice of Experimental Philosophy*, Peterborough, Broadview Press.

Sytsma, J. and Machery, E., 2009, "How to Study Folk Intuitions about Phenomenal Consciousness", *Philosophical Psychology* 22 (1), pp. 21–35,

Talbot, B., 2009, "Psychology and the Use of Intuitions in Philosophy", *Studia Philosphica Estonia*, 2.2, pp. 157–176.

Timmons, M., 2008, "Toward a Sentimentalist Deontology", in W. Sinnott-Armstrong (ed.) *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, Cambridge, MA, MIT Press, pp. 93–104.

Tobia, K., Chapman, G., and Stich, S., 2013, "Cleanliness is Next to Morality, Even For Philosophers", *Journal of Consciousness Studies*, 20, pp. 195–204.

Valdeso, P. and DeSteno, D., 2006, "Manipulations of Emotional Context Shape Moral Judgment", *Psychological Science*, 17, pp. 476–477.

Weinberg, J. and Alexander, J., 2014, "The Challenge of Sticking With Intuitions Through Thick and Thin", in A. Booth and D. Rowbottom (eds.), *Intuitions*, Oxford, Oxford University Press, pp. 187–212.

Weinberg, J., Nichols, S., and Stich, S., 2001, "Normativity and Epistemic Intuitions", *Philosophical Topics*, 29, pp. 429–460.

Weiskopf, D., 2009, "The Plurality of Concepts", *Synthese*, 169, pp. 145–173.

Wetherson, B., 2003, "What Good are Counterexamples?", *Philosophical Studies*, Vol. 115–1, pp. 1–31.

Sanja Srećković
Institute of Philosophy,
University of Belgrade

Original Scientific Paper
UDK 81'23:165
159.955.6.072:165

REASONING OF NON- AND PRE-LINGUISTIC CREATURES: HOW MUCH DO THE EXPERIMENTS TELL US?[1]

Abstract. *If a conclusion was reached that creatures without a language capability exhibit some form of a capability for logic, this would shed a new light on the relationship between logic, language, and thought. Recent experimental attempts to test whether some animals, as well as pre-linguistic human infants, are capable of exclusionary reasoning are taken to support exactly that conclusion. The paper discusses the analyses and conclusions of two such studies: Call's (2004) two cups task, and Mody and Carey's (2016) four cups task. My paper exposes hidden assumptions within these analyses, which enable the authors to settle on the explanation which assigns logical capabilities to the participants of the studies, as opposed to the explanations which do not. The paper then demonstrates that the competing explanations of the experimental results are theoretically underdeveloped, rendering them unclear in their predictions concerning the behavior of cognitive subjects, and thus difficult to distinguish by use of experiments. Additionally, it is questioned whether the explanations are rivals at all, i.e. whether they compete to explain the cognitive processes of the same level. The contribution of the paper is conceptual. Its aim is to clear up the concepts involved in these analyses, in order to avoid oversimplified or premature conclusions about the cognitive abilities of pre- and non-linguistic creatures. It is also meant to show that the theoretical space surrounding the issues involved might be much more diverse and unknown than many of these studies imply.*

Keywords: *cognitive processes, deduction, probabilistic reasoning, animal cognition, infant cognition*

Introduction

Theories in cognitive science and philosophy of mind strive to fit the results of experimental psychology. However, the results themselves typically do not hand us conclusive answers, so they have to be analyzed, and the analysis is in turn subject to theoretical assessment. In this paper I examine the assumptions behind the analysis of certain experimental results concerning the possible reasoning mechanisms of non- and pre-linguistic creatures.

[1] The work on this paper has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia through the project Logico-epistemological bases of science and metaphysics (No. 179067). I am also very grateful to Jacob Beck for reading the drafts of the paper and making useful comments about it.

doi: 10.5937/BPA1831115S

Certain behavior is consistent with a capability for reasoning from an excluded alternative: the recognition that if there are only two possibilities and it is not one of them, it must be the other. Recently, there have been attempts to experimentally test whether some animals, as well as pre-linguistic human infants, are capable of this kind of reasoning. I focus on the analysis of two studies which test the reasoning abilities of these creatures by setting up a task that needs to be solved. Call's (2004) two cups task tested whether great apes are capable of reasoning by exclusion, and since the apes proved to be successful at the task, the other study – Mody and Carey's (2016) four cups task – proposed a way to determine the mechanism responsible for that kind of reasoning.[2] I will demonstrate that there is a mistaken assumption in the analysis of the latter study, and that we cannot decide among the competing explanations based on the strategy proposed by the authors. I will further show that the two main competing explanations cannot be clearly distinguished from each other, because their assumptions, requirements, and implications are not sufficiently defined. As a consequence, the experiments conducted so far, as well as future attempts to decide between them using experimental research, might be misguided.

Reasoning by exclusion and possible underlying mechanisms

In Call's task, the subjects see the experimenter hiding the reward in one of two cups, not knowing which one. They then receive evidence that one cup is empty. If they reason by exclusion, they should use the information about the empty cup to exclude that location, and instead select the other cup. The subjects proved successful at this task (the rate of correct choices was significantly above chance), and their success can be interpreted and explained by several accounts. The term "reasoning by exclusion" does not make commitments as to the particular reasoning mechanism being used (Mody & Carey, 2016).

The results of this task were taken by many commenters to indicate that the animals manifest a capability for reasoning by the disjunctive syllogism. The subjects supposedly reasoned: The food is either in A or B. It is not in A. Therefore it is in B. This interpretation imposes the highest cognitive requirements: it requires the subjects to be representing the concepts OR and NOT as well as the dependent relationship between A and B (embodied

[2] The reader will notice that the two studies are done with different subjects (the latter study was done with human children). Moreover, the older children that were successful in the latter experiment clearly do not fall into the group of pre-linguistic creatures. Although this would make a difference for the plausibility of the thesis that is being tested, I, however, do not take this difference to be significant for my analysis, since I am interested in the methodology of these experiments, rather than in the results themselves. I focus on the difficulty of interpreting the behavioral data, regardless of the particular subjects and results of the study.

by the concept OR). The dependent relationship is what leads the subjects to update their representation of B when they see that A is empty, and to conclude that B necessarily contains the reward. Updating of B based on the information about A is referred to in the literature as 'inferential updating (Mody & Carey, 2016).

Another interpretation is named “Maybe A, maybe B”. The subjects initially believe that the reward might be in A, and also that it might be in B. The two possible hiding locations are regarded independently: gaining the information about A does not lead to inferential updating which would form a new appraisal of B. Thus, when they see that A is empty, they do not search in A anymore, and they proceed to B according to their initial premise that it *might* contain the reward (Mody & Carey, 2016).

The results can also be explained with even fewer cognitive demands. According to the “Avoid empty” interpretation, the animals have no particular beliefs about whether B contains the reward. They are merely not searching in A. When they see that A is empty, they avoid searching in it, and instead approach B merely because it is the other salient hiding location available. This does not even require representing the alternatives A and B as potential locations of the reward, so there is also no inferential updating (Mody & Carey, 2016).

Rescorla (2009) proposed another possible mechanism underlying exclusionary reasoning: probabilistic inference over the space of cognitive maps. Rescorla defines cognitive maps as mental representations that represent geometric features of the physical environment. Their most important feature is not having logical form. This allows them to be realized without the subject’s capability for logic. The key part of Rescorla’s analysis is the Bayesian decision theory, which gives calculations for the distribution of probabilities over cognitive maps. Probabilities assigned to cognitive maps can be understood as degrees of belief assigned to different hypotheses. The subjects recognize two possible locations of the hidden reward, represented by two cognitive maps (M1, M2). The maps correspond to two hypotheses concerning which cup has the reward in it. The subjects initially lack evidence regarding where the reward is hidden, so the initial probability distribution treats cups the same:

$$p(M_1) = p(M_2) = 0.5$$

Since they exhaust the space of possibilities, their probabilities sum up to 1.

Also, for each cup there are two possible outcomes: y_i – the cup has food in it, and n_i – the cup is empty, and their probabilities also sum up to 1: $p(y_i) + p(n_i) = 1$.

Assuming M_i is true, the probability that the subjects will recognize that the reward is in the cup is taken to be $p(y_i|M_i) = 0.8$, since the account allows the small possibility of the subject not seeing the reward even if it is looking in the right cup.

Since $p(y_i)+p(n_i)=1$, it follows that the chance of false negatives is $p(n_i|M_i)= 0.2$

There is also a slight chance of false positives, say, $p(y_i|M_j)=0.1$, which makes the chance of a correct observation that the cup is empty $p(n_i|M_j)=0.9$.

When the subject is presented with the evidence that the first cup is empty, the probabilities are updated over the cognitive maps using Bayes' Law:

$$p(M_1|n_1) = 0.182$$
$$p(M_2|n_1) = 0.818$$

The initial probability of 0.5 is lowered for M_1 to 0.182, while M_2 is updated to 0.818, by the process of inferential updating (since the two cups are represented as being in a disjunctive relationship). Thus, the subject reaches a conclusion that it is more probable that the reward is in cup B than that it is in A.

All four interpretations can explain the experimental results, because using any of these approaches would lead subjects to be successful in the task. Subsequent experiments were designed to distinguish between these interpretations. I focus on Mody and Carey's (2016) study, which tested for behavioral evidence of inferential updating. This should then distinguish between the deductive and probabilistic interpretations on one side, and the remaining two interpretations which predict no inferential updating on the other side.

Distinguishing between the interpretations

The following experiment was conceived as an extended version of Call's task. Two rewards (in this study the rewards were stickers) were hidden in four cups, one reward in each of two pairs. The first pair of cups was covered by a screen so that the subject could not see which of the cups the reward was placed in, and then the same was done with the second pair. The participants were children from 2.5 to 5 years old, divided into four groups by their age. When one of the cups was revealed to be empty, the child was supposed to pick the other cup from the pair, which is certain to contain a reward.

If children were using the deductive syllogism, they would engage in inferential updating, meaning that the information about the empty cup (not-A), in combination with the representation of where the sticker was hidden (A or B), would lead them to conclude that the other cup from the pair necessarily contained the sticker (therefore B). This interpretation predicts children will choose the "target cup" (B).

If they were using the "Maybe A, maybe B" strategy, obtaining the information "not-A" would not lead to updating information about B. The children would still hold on to "Maybe B", and all the remaining cups would seem to be equally good candidates. Thus, the children would choose the target cup at an equal rate as the other two cups.

According to the “Avoid empty” strategy, learning “not-A” would lead to avoiding A, but without representations about other possible locations. Thus, the subjects would also be expected to choose all three remaining cups at an equal rate (Mody & Carey, 2016).

Mody and Carey seem to take the probabilistic account to predict that the children will pick the target cup preferentially to the other cups (due to inferential updating). Still, they seem to take probabilistic reasoning as somewhat “less certain” than deductive reasoning. Thus, the probabilistic interpretation might predict preferential choosing of the target cup, but at a somewhat lower rate than in the deductive scenario. I will address this in more detail in the following section.

Prior to the main task, there was a training phase, which involved only three cups: one pair of cups and one single cup, hidden behind two screens. The procedure of hiding the rewards was the same: the first reward was placed in the single cup, and the second reward was placed in one of the cups from a pair. After removing the screen, the child was asked to choose a cup. This task did not require reasoning by exclusion, but it still required comparing the sure cup to the two uncertain cups.

Results and analysis

In the training trials chance was established at 33%, since there were three cups to choose among. The children chose the target cup at rates significantly above chance. In the test trials, since the children virtually never chose the empty cup, it was taken that three cups were the remaining options, and chance was also set at 33%. The results were very similar to the training phase. All groups except for the youngest chose the target cup significantly above chance, suggesting that they engaged in inferential updating. The youngest children chose the target cup in only 36% of the cases, not significantly above chance. They behaved in a manner predicted by the “Maybe A, maybe B”, and “Avoid empty” interpretations. The rates of choosing the correct cup in both phases of the experiment are given in the table.

Training trials		Test trials	
Age group	Success rate	Age group	Success rate
2.5	47%	2.5	36%
3	60%	3	58%
4	71%	4	64%
5	72%	5	76%

These results indicate that, except for 2.5-year-olds, the children chose the target cup preferentially, and thus behaved in a manner consistent with inferential updating. This allows accepting only the probabilistic and deductive interpretations as possible, while dismissing the others.

We can sum up the main issue as follows.

In the training trials, an observation that cup A is full leads to representing cup A as sure and cups B and C as unsure, which leads to reaching for cup A.

In the test trials, an observation that cup A is empty leads to representing cup A as sure (empty), cups C and D as unsure, and leads to updating the representation of cup B from unsure to sure (full), and reaching for it.

The two possible interpretations of the results aim to answer the question: What is the cognitive process that leads to updating of cup B from "unsure" to "sure (full)"?

According to the probabilistic interpretation, the subjects engage in Bayesian redistribution of *coupled* probabilities (synchronously lowering the probability of cup A and rising the probability of cup B), while representing cups C and D as *independent* probabilities and as remaining unsure.

According to the deductive interpretation, the subjects engage in a logical inference: "The reward is either in A or B. It is not in A. Therefore it is in B". Cups C and D are not included in the inference, since they are represented as independent from A and B.

Mody and Carey go further to claim that the deductive interpretation is more plausible. Their strategy was to additionally formulate the difference between these two cognitive processes in terms of *certainty*: deductive reasoning would lead to a choice based on *absolute* certainty that the reward is in cup B, while the probabilistic one would lead to a choice based on *increased* certainty (the subjects are only *more* certain that the reward is in B than that it is in any of the other cups). They then propose a way of distinguishing between these two mechanisms, claiming that one feature of the gathered data indicates absolute certainty behind the children's choice. Namely, the children chose the target cup just as often in test trials as in training trials – in which they could directly observe (and thus be absolutely certain about) where the sticker was hidden. These results suggest that children were absolutely certain in the test trials, too. In other words, since the rate of choosing the correct cup was the same in the trials which required reasoning as in the trials which did not, their reasoning was interpreted as absolutely certain, and therefore, deductive:

> Our design did not allow us to distinguish between a choice based on absolute certainty and one based on increased certainty. The latter would still require that children represented the dependent relationship between the two locations, and that they inferentially updated their assessment (...); however, the inference children made would not be truly deductive. This possibility was put forth by Rescorla (2009), who described it in a Bayesian framework, where the probability associated with one possibility is adjusted up as the probability of another possibility goes down. However, one feature of our data suggests that children were making a deductive inference: 3- to 5-year-old children chose the target cup just as often in test trials as they did in training trials, in which they could directly observe that a sticker was being hidden there (Mody & Carey, 2016, p. 46).

I will proceed to criticize Mody and Carey's characterization of the two reasoning mechanisms by different degrees of certainty. I will first demonstrate that this characterization is false, and then I will show that probabilistic updating of coupled probabilities cannot be distinguished from explicit inferences (for now), and that these two cognitive processes may even be only variants of each other, instead of being independent strategies.

Analysis of the analysis

Mody and Carey make a mistake of confounding two possible applications of the property of certainty. One is the certainty that defines deduction, and applies to the *transition* from the premises to the conclusion: in a valid deductive argument, the truth of the premises guarantees the truth of the conclusion. Thus, the subject can be certain that if the premises are true, the conclusion must also be true. But this does not tell us how certain the subject was in either of the premises, nor of the conclusion. The latter type of certainty applies to the propositions themselves, and it can have various degrees. The rules of "probability preservation", or "uncertainty propagation" are defined within propositional probability logics. The main idea is that the premises of a valid argument can be uncertain, in which case the conclusion will also be uncertain (Demey et al., 2017). Therefore, the deductive account does not necessarily predict children will be absolutely certain that the reward is in cup B. It needs to additionally postulate that they are absolutely certain of the premises. This type of certainty is what is of interest for the experiment, because the degree of certainty about the final conclusion is what affects the subjects' behavior, and thus the percentages that Mody and Carey appeal to. Let us see how this type of certainty is accommodated within the two accounts.

In the probabilistic account, the distribution of degrees of certainty is determined mathematically, according to the Bayes' Law. We saw in Rescorla's account that it allows the possibility of subjects making both false negative, and false positive judgements, due to their fallibility. This is reflected in not assigning absolute certainty to even the seemingly obvious observations, such as "A is empty." Thus, the "premise" ("Reward is in A") gains a probability slightly higher than 0. This, in turn, renders the probability of the conclusion "Reward is in B" as slightly less than 1. The probabilities assigned to cups C and D should be equal, and the same as in the initial distribution (0.5). Therefore, if the subjects reasoned probabilistically, they would indeed make a choice based on "increased certainty" of B over other options, just as Mody and Carey suggest.

In the deductive account, however, degrees of certainty have not been mentioned. The experimenters expect absolute certainty by default, and the percentages of failure of subjects to perform the task (which were at least 30% of choices, as we saw in the table) are explained by appealing to "noise" or "performance issues", such as limitations of attention, working memory, or other factors. (Mody & Carey, 2016, pp. 46–7). So far, it seems that the probabilistic account has a better formal apparatus for dealing with the degrees of certainty. Still, there are probably several ways in which they could be incorporated in the deductive account as well. One way would be to assign probabilities to the propositions, like it is done in the probabilistic semantics. We can have a probability function for the propositional language L, and the valuations v: $L \rightarrow \{0,1\}$ of classical propositional logic can be replaced with probability functions P: $L \rightarrow \mathbb{R}$, which take values in the real unit interval [0,1]. The classical truth values of true (1) and false (0) can thus be regarded as the endpoints of the unit interval [0,1]. This would mean taking classical logic as a special case of probability logic, or equivalently, taking probability logic as an extension of classical logic (Demey et al., 2017). Applying this to the cups task, we can formally express the subject's deductive reasoning as follows:

$P(A \vee B)=1$

$P(\neg A)=0.9$

Therefore, $P(B)=0.9$[3]

In cognitive terms, the probabilities could be defined within a meta-cognitive level, without being explicitly represented by the subject. Thus, even though the subjects reason through a logical inference, each step in the inference (e.g. $\neg A$) may be accompanied by a degree of subject's certainty about the step. However, it is difficult to specify how the probabilities of two or more premises are to be combined. Some advocates of this kind of extension of classical logic propose a rule that "a p-valid inference cannot take us from low uncertainty in the premises to high uncertainty in the conclusion". They define the uncertainty of a proposition p as one minus its probability, 1—P(p). Then an inference with two or more premises is p-valid if and only if the uncertainty of its conclusion is not greater than the sum of the uncertainties of its premises for all coherent probability assignments (Evans et al., 2015).

This version of the deductive account needs to be further theoretically developed. Nevertheless, the outline shows a way to implement different degrees of certainty into the deductive account. In application to Mody and Carey's results, even though it was shown that there are two separate applications of certainty, that still does not prove that they were wrong in assigning absolute certainty to subjects reasoning deductively. Indeed, we do not know how certain the subjects were of any of their propositions.

Difference between the competing accounts

This brings me to the final point of this paper. What *is* the difference between the deductive and probabilistic accounts? How can we behaviorally

[3] The numerical values of probabilities are just an example.

distinguish which of these two reasoning mechanisms the subject is using? I claim that the two accounts are not sufficiently developed, and not clear in their theoretical requirements. This renders them unclear in their predictions concerning the behavior of cognitive subjects, and thus difficult to distinguish by use of experiments. I will show this by presenting several possible candidates by which these accounts might be differentiated.

Degrees of certainty

As I demonstrated, since probabilities (and thus the degrees of certainty) can be accommodated within the deductive account, it is unclear whether the accounts differ in the degrees of certainty assigned to the conclusion. Thus, we do not know whether it is possible to differentiate them empirically – whether they predict different percentages of successful task performances. It is yet to be shown that there would be a difference at all.

Format of mental representations

Another way to distinguish them might be by the format of the mental representations they posit. The accounts were presented as positing different kinds of mental representations: the probabilistic reasoning was presented as defined over *cognitive maps*, while the deductive reasoning is taken to be computed over *proposition-like* mental representations, and made available by language (Bermudez, 2006). Rescorla's probabilistic account was formulated partly as a way to enable computing sophisticated reasoning over non-propositional mental representations. However, neither of these accounts is necessarily tied solely to their respective representational formats. Probabilistic reasoning may also be computed over propositions – the hypotheses which the probabilities are assigned to may as well be in the form of propositions. That is, in fact, exactly how the probability distribution over competing hypotheses is most often presented (Rescorla, 2009).

In addition, even though it is not the most popular opinion among cognitive scientists, there are some authors (e.g. the advocates of diagrammatic reasoning) who claim that logical reasoning can be defined over non-linguistic representations, and that there is no intrinsic difference between symbolic and diagrammatic systems as far as their logical status goes (Shin & Lemon, 2018). This would imply that proposition-like mental representations might not be necessary for logical reasoning. Thus, each of these two accounts could be modelled quite differently from the versions of them proposed so far, and this would certainly reflect on their behavioral implications, significant for the experimental testing.

Logical structure

The most important difference between these accounts is supposed to be whether they commit to logical structure – whether they describe the subject's reasoning as proceeding by logical rules, or by some other kind of inference. The deductive account clearly appeals to the logical structure of the deductive syllogism. The probabilistic account purportedly does not involve a logical structure, but is instead structured as a distribution of probabilities over a space of hypotheses (which in Rescorla's account have the form of cognitive maps). However, this attempt at differentiating the two accounts also has its difficulties. First, it is not clear how this difference might, if at all, be behaviorally manifested. Thus far, we have no means to experimentally test between these accounts. Second, it is an open question whether Bayesian reasoning is truly an alternative to reasoning by the disjunctive syllogism, or one way of implementing it – which might explain why they are also difficult to differentiate behaviorally. As Mody (2016) observed,

> the construction of the hypothesis space [in the probabilistic account] requires that children enumerate the relevant possibilities, and the inference mechanics maintain a fundamentally disjunctive relation between them. Further, the lowering of probability associated with gaining negative information essentially implements negation. Thus, even if reasoning proceeds probabilistically, propositional representations including negation and disjunction might be required to represent the information that the probabilistic mechanism uses as input.

In other words, it is unclear whether probabilistic reasoning is, in fact, dependent on some form of logical reasoning, or on some logical concepts at least. It is at least sophisticated enough to involve the ability to distinguish between coupled and independent probabilities. I agree with Mody that the experimental results presented here do give evidence of representing negation and disjunction in some way. However, they do not necessarily imply full-blown logical inference.

Simplicity of explanation

In referring to the results of similar experiments indicative of the existence of basic logical reasoning in 12- and 19-month-old human infants, Arlotti and colleagues (2018) admit that a Bayesian probabilistic model of reasoning could also explain the results: "Bayesian iterative models (...) could mimic deductive syllogism without assuming a logical inference." However, they argue that the probabilistic explanation, although compatible with the results, has more requirements than if we only assume the infants are performing a logical inference. It requires that infants represent the space of alternatives (which is equivalent to implementing a disjunctive representation), assign ordered priors to the alternatives, and assess alternative evaluations iteratively (Arlotti 2018).

Arlotti and colleagues thus defend the deductive explanation of their own results, but differently than the previous authors – merely as a more parsimonious explanation of the results. However, this defence is still not unquestionable. One of the reasons is that it is not clear whether the probabilistic account actually commits to the assumptions brought up by Arlotti, especially if those assumptions imply an explicit representation of assigning probabilities to hypotheses, or of assessing the alternatives. Rescorla (2009) describes this assignment of probabilities as consisting in "suitable function relations between cognitive maps and mental representations denoting numbers". For a cognitive subject to assign probabilities really means "to enter a mental state bearing appropriate functional relations to other mental states". Rescorla seems to think that the complex probabilistic computations can be performed at lower-level processes of cognition (and perception), and does not really clarify what exactly the probabilistic account posits as being explicitly represented by the cognitive subjects. Thus, the last possible candidate used for deciding between the deductive and the probabilistic accounts – parsimoniousness of the explanation – is also rendered unusable, due to the lack of knowledge about the exact assumptions of the accounts.

In conclusion, it is difficult to distinguish the reasoning mechanisms by their behavioral signatures when the implications of the theories that posit them have not been made clear. These accounts have not been developed in sufficient detail, and the experimental psychologists should bare this in mind in order to avoid oversimplified or premature conclusions about the cognitive abilities of pre- and non-linguistic creatures. In addition, the theoretical space surrounding these issues might be much more diverse and unknown than these studies imply.

References

Bermúdez, José Luis (2006). Animal reasoning and proto-logic. In S. Hurley and M. Nudds (eds.), *Rational Animals?* Oxford: Oxford University Press, 127–138.

Call, Josep (2004). Inferences about the location of food in the great apes (Pan paniscus, Pan troglodytes, Gorilla gorilla, and Pongo pygmaeus). *Journal of Comparative Psychology*, 117–128.

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., Bonatti, L. L., (2018) Precursors of logical reasoning in preverbal human infants, *Science* 359, 1263–1266.

Demey, Lorenz, Kooi, Barteld and Sack, Joshua (2017). Logic and Probability, *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), Retrieved from https://plato.stanford.edu/archives/sum2017/entries/logic-probability/

Evans, J. St. B. T., Thompson V. A. and Over D. E. (2015). Uncertain deduction and conditional reasoning. *Front. Psychol.* 6, 398.

Mody, Shilpa & Carey, Susan (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, 154, 40–48.

Mody, Shilpa (2016). *The Developmental Origins of Logical Inference: Deduction and Domain-Generality*. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Rescorla, Michael (2009). Chrysippus' dog as a case study in non-linguistic cognition. In Robert W. Lurz (ed.), *The Philosophy of Animal Minds*, Cambridge University Press, 52–71.

Shin, Sun-Joo, Lemon, Oliver and Mumma, John (2018). Diagrams, *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), Retreived from https://plato.stanford.edu/archives/sum2018/entries/diagrams/.

Sabrina Coninx
Albert Newen,
Ruhr-Universität Bochum[1]

Original Scientific Paper
UDK 165.242.2

THEORIES OF UNDERSTANDING OTHERS: THE NEED FOR A NEW ACCOUNT AND THE GUIDING ROLE OF THE PERSON MODEL THEORY

Abstract. *What would be an adequate theory of social understanding? In the last decade, the philosophical debate has focused on Theory Theory, Simulation Theory and Interaction Theory as the three possible candidates. In the following, we look carefully at each of these and describe its main advantages and disadvantages. Based on this critical analysis, we formulate the need for a new account of social understanding. We propose the Person Model Theory as an independent new account which has greater explanatory power compared to the existing theories.*

1. Introduction

Humans are hyper-social beings that are highly dependent on adequate interaction with others. Right after birth our survival depends on social interaction, and this remains a key aspect of biological, economic and social success throughout our entire lives. Given the phylogenetic and ontogenetic relevance of social interaction, researchers across disciplines aim for an adequate theory of how humans are able to understand others. This aim has not yet been definitively reached, but recently the debate has received new input after a decade of stagnation (in the 90s) within a philosophical debate virtually restricted to the choice between Simulation Theory and Theory Theory. Now, multiple accounts are on the table. This motivates us to clarify the main positions, their arguments and their relations to each other. We proceed as follows: first, we dedicate one section apiece to each central (class of) positions, namely Theory Theory, Simulation Theory and Interaction Theory. Based on the advantages and disadvantages that these accounts reveal, we argue in a second step in favor of what we call Person Model Theory.

[1] For fruitful contributions we would like to thank Atesh Koul and Julia Reh as well as two anonymous reviewers. The paper was supported by the DFG-Research Training Group (no. GRK2185) "Situated Cognition" and by the DFG-project (NE 576/14-1) "The structure and development of understanding actions and reasons".

doi: 10.5937/BPA1831127C

2. Philosophical Theories of Mind-Reading

Three competing alternatives enjoy the lion's share of discussion within the philosophy of 'mind-reading'. Theory Theory introduces theory-based inferences relying on folk-psychological rules as the central basis for understanding others. This view may account for a great deal of explicit and reflective social understanding after age 4 or 5. However, it seems to be inadequate to account for the intuitive understanding of others which develops rather early in infancy and remains active even after we have acquired the capacity for explicit theory-based inferences. Simulation Theory is especially suited to account for the early intuitive social understanding that takes place on the basis of simulating the mental state of other subjects. However, one main deficit is that the simulation process may not be possible in many real-life situations, including observations of persons with mental disorders or radically different cultural backgrounds. Moreover, it appears questionable whether this theory is an adequate framework to integrate the basic mirror neuron story on which most defenders of this account strongly rely. Interaction Theory argues plausibly for the importance of direct and smart perception in social understanding as well as for the distinguishing role of online interaction. However, when seen in the light of recently available empirical evidence, Interaction Theory overestimates the primacy of basic forms of direct social coordination.

2.1 Theory Theory

The core idea of "Theory Theory' (TT) is the claim that the capacity to understand others is based on a folk-psychological theory that is used for systematic inferences. Humans employ a folk-psychological theory (i.e. an abstract and coherent system of law-like assumptions) to derive the mental states of others, such as their beliefs, attitudes, desires or emotions, and to thereby anticipate their future behavior.

A modular version of TT is provided by Baron-Cohen (1995), according to which the human psychological system is composed of various modules which interpret the world in accordance with an inborn organizational structure evolved through natural selection. Each module is tailored to solve a certain adaptive problem, and correspondingly there exists a specialized mindreading system designed to comprehend and predict the behavior of other subjects. Due to its linkage with perceptual processes, this system enables the recognition of visible cues reliably indicating the internal mental states of others. The flexible and fast inference of the complete range of mental states from observable signals depends on the management of the 'Theory-of-Mind-Mechanism' (ToMM).

In contrast, Gopnik & Wellman (1992) assume that children gradually develop a Theory-of-Mind (ToM) ability that is based on the same cognitive mechanisms that adults apply in the development of scientific theories. During a constant learning process, children generate general assumptions about unobservable entities, formulate expectations, adjust their theory in accordance with the evidential data they collect through experience, invent auxiliary hypothesis and replace their theoretical constructs and rules in the light of continuously occurring counter-evidences. The progressive improvement of this causal-explanatory theory about how others come to perform specific actions finally leads to a coherent representational system of propositional attitudes.

Accordingly, the main difference between those two accounts concerns the acquisition of the ability to explicitly represent the mental states of others. Gopnik & Wellman claim that the ToM ability is based on a psychological theory which passes through the same dynamic process of prediction, falsification and adjustment as do scientific theories, until it reaches the status of a mature theory which complies with the demands of a full-blown ToM ability. Conversely, Baron-Cohen argues for a phased maturing of distinct innate modules, where the ToMM comes into operation only in the final stage. Nevertheless, both accounts agree on the assumption that from a certain point in infantile development humans refer to a complex theory-like structure of mentalistic knowledge to infer the propositional attitudes of others.

This idea is supported by experimental studies proving that approximately around the age of 4 children possess the mentalistic abilities to pass the so-called 'false belief task', an experimental setup which has been implemented in widely known versions by Wimmer & Perner (1983). According to Baron-Cohen (1995), at this age children have mastered the use of the ToMM, while Gopnik & Wellman (1992) interprets the results as a further improvement of their causal-explanatory theory based on previous experiences of false prediction. Independently from the question of ontogenetic genesis, the employment of theory-based inferences is treated as the general epistemic strategy used by older children and adults in everyday life. This inferential mindreading mechanism proves especially useful in understanding other subjects which differ fundamentally in their mindset and their behavioral patterns from oneself. Such situations might occur in contact with members of other cultures, with persons suffering from mental diseases, or with animals (Newen 2015a). Furthermore, humans tend to use the epistemic strategy of TT when a social situation offers merely a small number of perceptual cues for the internal state of persons involved (Baron-Cohen 1995).

According to TT, theory-based inferences are the primary mentalistic strategy. However, TT overlooks the possibility of having direct access to many basic mental phenomena, simply by simulating the other's situation or directly perceiving their mind state. It overintellectualizes intuitive understanding in early infancy and it underestimates the role of second-person involvement as well as the role of one's own experiences in understanding others.

The TT assumes the employment of a third-person viewpoint towards another person's mental states in a manner analogous to scientific inquiry. Nevertheless, the ability to comprehend and predict behavior in mentalistic terms becomes particularly important in interpersonal cooperation where the mindreading person does not merely function as an observer but as a dynamic interacting part (Di Paolo & De Jaegher 2012). The worry is that the observational stance which is usually adequate in science is only an exceptional perspective in understanding others, while humans are frequently involved in second-person interactions (Vogeley, Schilbach & Newen 2013; Schilbach et al. 2013). Furthermore, instead of relying on a theory, people sometimes just rely on their own sparse experiences in similar situations as a basis for mindreading (Goldman 1992).

The method of the TT is grounded on the assumption that internal mental states are only accessible due to complex inferences whereas observational behavior constitutes the evidential basis for further theoretical considerations. Conversely, understanding others does not always require such intellectual capacities (Gallagher 2008). At least within human culture, we seem to possess the universal ability to directly perceive basic emotions (Ekman & Friesen 1971; Gallagher 2008; Newen et al. 2015). Even young infants are able to intuitively understand others although they have not yet acquired an explicit or implicit theory of systematically interconnected beliefs. Ontogenetic studies clearly demonstrate that infants of less than one year of age are sensitive in their reactions to the affective expressions of caretakers, as in the visual cliff experiment (Sorce et. al. 1985), and they expect a smooth interaction pattern which leads to irritation if not used, as in the still face paradigm (Weinberg et. al. 2008).

2.2 Simulation Theory

In contrast to TT, 'Simulation Theory' (ST) dismisses the assumption that humans use a specific 'theory' to understand other people's minds. Rather, subjects simulate the others' situations and 'put themselves in the other's shoes'. The ST account does not need to presuppose a generalized set of laws similar to science, and it is characterized as information-poor mind-reading (Goldman 1992) as it does not presuppose an interconnected set of beliefs or belief-like information. However, as argued by Gordon (1986, 1992), simulating other minds does not merely mean to project one's own situation, as it also requires necessary adjustments concerning other persons and their perspectives. It is suggested that ST is routed in phylogenetically and ontogenetically basic mechanisms, taking advantage, for instance, of human abilities to read gaze direction or to imitate others (Gallese & Goldman 1998). Simulation enables subjects to generate explanations for the behavior of others and to predict how they are most likely to act in the future (Goldman 1989; Gordon 1992; Spaulding 2010). Some of these simulations are thought of as conscious and voluntary processes (Goldman 2006), others as unconscious and automatic in the sense that they do not require access or control over the stimulation processes (Gordon 1992).

Egocentric errors or biases, i.e. the influences of the mindreader's own mental states on the ascription of mental states to others, provide the first evidence for ST (Goldman & Jordan 2013). A paradigmatic example is the so-called curse of knowledge. This term designates the phenomenon that the participant's own knowledge influences his attribution to another person, although the participant is informed about the difference between their own and the other person's knowledge (Birch & Bloom 2003; Camerer, Loewenstein & Weber 1989; Nickerson 1999). ST also gathers support from neurophysiological studies where, for instance, amygdala lesions do not only strongly reduce the experience of fear in patients, but also their ability to recognize the fear of others on the basis of their facial expressions (Adolphs et al. 1994).

The interdependence of first-person experience and third-person observation receives further evidence from the discovery of mirror neurons. The class of mirror neurons, first discovered in monkeys, is active both during the performance of an action as well as during the observation of another individual performing this very action (Di Pellegrino et al. 1992; Rizzolatti & Craighero 2004). It has been proposed that when we observe someone perform an action, activation in our mirror neuron system simulates the action 'as if' we were performing it. The discovery of mirror neurons is supposed to be the most striking evidence for ST, as they are supposed to constitute the neural realization of at least the automatic forms of simulation. The mirror neuron system has thus been proposed as the basis for our understanding of others (Gallese & Goldman 1998; Keysers & Gazzola 2009; Rizzolatti & Craighero 2004; Sinigaglia & Rizzolatti 2008).

Despite the supporting evidences for ST, the theory faces several main issues. First, one can think of many cases in which subjects reliably predict the experience and behavior of others without being able to simulate them. For instance, ST is not necessary to understand persons with mental disorders, such as patients suffering from delusion of persecution, or persons who exhibit idiosyncratic, irrational behavior (Tversky & Kahneman 1974). The same also holds for beings involved in radically different cultures (Newen & Schlicht 2009). Their minds are simply too different from one's own to apply the epistemic strategy of simulation. Nonetheless, persons who possess general or specific knowledge about the respective subjects are able to understand what is going on in their minds and to successfully interact with them (Newen 2015a). This general or specific information we make use of can be learned as rules-of-thumb or an explicit theory, e.g. how to deal with a schizophrenic family member, without being able to simulate this person. Understanding based on behavioral rules-of-thumb or a theory can be quite advanced and enables smooth interaction despite lacking short-term or even long-term simulating or intuitive access to the deviating mindset of the others. To justify this relevance of rules-of-thumb or explicit theories in cases of mental disorders, we appeal to everyday experiences in dealing with persons with different mindsets (and without knowing this mindset), but we can also rely on studies with Asperger autistic people: they have a large deficit in all types of intuitive or simulative epistemic access to others but they can learn to partially compensate by learning to apply explicit theoretical rules. This indicates that simulation is not necessary and that non-autistic people rely on a plurality of epistemic strategies, not only simulation which can be lacking (for details see section 3.2 the pathology argument).

Second, ST remains limited in the sense that it adopts a first-person perspective in which the simulating individual is still considered as an observer (Gallagher 2008). Our requirements for understanding others' actions, however, is critical when we are interacting with them (Schilbach et al. 2013) whereby this online interaction is often realized with non-simulative but complementary actions (de Bruin et al. 2012). Third, the discovery of mirror neurons does not so far explain the relation between the first-person and third-person perspective. Mirror-neurons encode for certain types of actions and emotion, but they do not provide an answer to the question of how we attribute internal states to others on the basis of these neural processes. Moreover, the neural correlate in the case of third-person attribution of, for instance, beliefs does not involve the most characteristic correlates of first-person attribution of belief (Vogeley et al. 2001; Vogeley & Newen 2002), whereas ST would expect such an involvement. Thus, despite the important discovery of the mirror neuron system, its role in the process of understanding others still needs to be worked out in detail and its function in cases of simulation (which might sometimes happen) needs to be complemented by further neural processes. As long as this part of the story is missing, the mirror neuron system remains an interesting and still important component for automatic social processing (Neufeld et al. 2016), but this component still needs to be integrated into a theory of understanding others.

2.3 Interaction Theory

Interaction Theory (IT) is a phenomenologically inspired approach which claims that we understand others primarily and most importantly in situations of direct social interaction, which leads to the distinction between *online* and *offline* forms of social understanding (Frith & Frith 2003). More precisely, IT combines at least two claims: one about the important role of direct perception of mental states of others independent from any inferences (Gallagher 2008), and one about the primacy of understanding by adequate interaction (Hutto & Gallagher 2008).

IT characterizes human 'mindreading' as a form of smart perception. This means that the content of our perceptual experience can be rich and include mental phenomena in the sense that we can directly perceive the internal states of other subjects. While some argue that the contents of perceptual experiences are exclusively low-level properties (Tye 1995), in recent years many people have argued that the contents of perceptual experiences can also involve high-level entities such as causal relations (Butterfill 2009; Siegel 2009), actions and agency (Gao et al. 2009; Rutherford & Kuhlmeier 2013). In the same way, a phenomenological perspective is often used to argue for the rich content of our perceptual experience in social cognition, prominently defended by Gallagher (2008) and Zahavi (2011). The general line of argument can be roughly characterized as follows: perceptual experiences can be cognitively penetrated and they can thereby involve a rich content (McPherson 2012; Vetter & Newen 2014; Newen & Vetter 2016). Expert perception, we may say, is different from the perception of laypersons. A chess expert has a richer perceptual content when looking at a chessboard compared to a novice (Newen 2017). Since humans are hyper-social beings and, thus, most likely experts in understanding others, we are able to have a rich content in our social perception, e.g. in the perception of others' emotions (Zahavi 2011; Marchi & Newen 2015) or intentions (Pacherie 2005).

The relevance of direct perception has been convincingly argued for, and as a consequence even some representatives of TT have recently started to include direct perception as an important epistemic tool (Herschbach 2012; Carruthers 2015). Thus, direct perception appears as a certain kind of epistemic strategy that might be employed in different forms of mind-reading, even from a third-person perspective. In contrast, IT accounts rely in large parts on the assumption that the central constituent of understanding others is direct perception in online interaction which highlights the relevance of the second-person perspective in mind-reading (Gallagher 2002, 2008). Different versions of IT allow for several strategies of understanding others, all of which assume the primacy of understanding by interaction (Hutto & Gallagher 2008). De Jaegher & Di Paolo (2007), for instance, claim that the constitutive feature of all cases of online interaction is participatory sense-making where this is explained in terms of coordination. According to them, the process of coordination in interaction is constitutive in many cases of social understanding. The main examples to support this claim are cases of special joint action based on mutual social understanding, such as ballroom dancing.

Some evidence for the relevance of social interaction for social understanding is drawn from developmental psychology, which distinguishes the capacity for primary, secondary and sometimes in addition tertiary intersubjectivity (Trevarthen & Hubley 1978; Trevarthen 1979). Primary intersubjectivity involves the ability to reciprocate in face-to-face exchange and starts from two months of age onwards and thereby goes beyond the very early pure imitation abilities. It is, for instance, demonstrated in the still-face-paradigm (Bertin & Striano 2006). Consequently, we have a communication basis that allows even infants to exchange and read common cues via bodily movements, gestures, facial expressions, eye direction, etc. Secondary intersubjectivity is typically realized when triadic intentional communication begins, e.g. in interactions involving joint attention which start at approximately 9 months. This secondary level involves the understanding of other people while acting together in a pragmatic context. It permits sharing and coordinating with another person's attention, feelings and intentions toward a third object, event, or action (Trevarthen & Hubley 1978). While it is assumed that primary intersubjectivity is innate and allows even newborns to perceive other person's mental processes, secondary intersubjectivity develops later in the first year of life. Tertiary intersubjectivity develops when children aged 4 begin to employ an ethical stance by beginning to manifest explicit rationale about what is right and wrong, as well as explicit attitudes about others' mental states (Trevarthen 2006).

In addition, there is now more and more evidence that social cognition is fundamentally different when we are in interaction with others rather than merely observing them. This is shown by systematic investigations of the underlying neural processing, e.g. in a test of observing facial expressions which are either directed towards oneself or towards another. While self-directed facial expressions lead to a differential increase of neural activity in the ventral portion of the medial prefrontal cortex and the (superficial) amygdala, other-directed facial expressions result in a differential recruitment of medial and lateral parietal cortex (Schilbach et al. 2006). In another study (Schilbach et al. 2010) of two persons either realizing joint attention towards an object or looking at different objects, it was shown that joint attention had a specific neural profile which closely matches with the so-called mentalizing network relying on the medial prefrontal cortex and posterior cingulate cortex. Furthermore, it was shown that producing joint attention (e.g. directing someone else's gaze toward an object) activated the ventral striatum (i.e. reward system). This indicates that activating joint attention is pleasurable for healthy people. Together with other evidence, this triggered the claim that we should presuppose a *second-person neuroscience* (Schilbach et al. 2013) and, thus, it is convincingly argued that understanding others in a situation of online-understanding is systematically different from understanding others by observation without interaction.

The most important insight is delivered by the claims that online-understanding is a specific form of understanding in contrast to offline-understanding, and that direct perception plays a decisive role in social understanding. This is the strongest feature of IT, but it still leaves us with the open question whether online understanding is in fact primary to offline understanding, and if so in which sense – phylogenetically, ontogenetically or even constitutively. It remains questionable to what extent observations of simple coordination can be generalized to all cases of social understanding and whether it is prior in comparison to the diverse other forms of social understanding (Andrews 2012; Newen 2015a). The evidence here is uncertain but evolutionary considerations may speak for the claim that both are equally relevant strategies of understanding. To survive as social beings, humans need to learn from both interaction and observation as much and as soon as possible. Thus, a primacy claim leaves the burden of proof on the side of IT.

Furthermore, IT overlooks the importance of the long-term social relationships which are habitual and re-activated in social interactions, e.g. in the case of understanding a familiar person. This long-term person-centred information can become strongly relevant in shaping an online interaction, much more so than any specific information about the situation in which one deals with this person, and it is also relevant in offline understanding, such as when trying to understand the familiar person while discussing him or her with a friend. This criticism can be condensed into one core difference for which IT cannot account: namely, the difference between the social understanding of a person's actions in one and the same situation type, where in one case the person is a complete stranger and in the other a well-known person such as a family member or a friend. This is especially relevant since this difference is already implemented in early infancy, e.g. the phenomenon of infant shyness in which infants react shyly to adult strangers, which manifests during the third quarter of the first year.[2] The relevance of prior information in the evaluation of a person's mind-set is also reflected in empirical studies investigating the impact of stereotypes. Culturally anchored stereotypes (Macrae & Bodenhausen 2000) and stereotypes in general (Macrae & Quadflieg 2010) substantially shape our understanding of others (review: Newen 2015a, sec. 5.1–5.3).

3. Person Model Theory

3.1. Definition

The central idea of the 'Person Model Theory' (PMT) is twofold: On the one hand, we need to accept that humans use *a multiplicity of epistemic strategies* (theory-based inferences, simulation, direct perception, contextual or narrative embedding) to account for all cases of understanding others. On the other hand, we need to take into account that humans rely on prior information stored in form of *person models* and *situation models*. As such,

[2] Defenders of IT may reply that they can include *memorized* interactions to account for these facts. However, this would require substantial alterations of the innate proposal of IT in accepting memorized models of other persons. Another move would be to claim that the memorized information is available in the form of narratives, since those are an additional tool in IT. Although narratives are an important instrument to enrich information about others which unfolds from 2 years of age onwards (Hutto 2008; Newen 2015), they cannot account for the relevant sensitivity in early infancy.

the PMT according to Newen (2015a, 2017) aims to answer two mainly independent questions.[3]

The first question asks which epistemic strategy humans use to access the mental states of others and to gather information about them. Concerning the *epistemic strategy*, PMT defends the multiplicity view: we do not rely on one epistemic strategy as is suggested by most proposals in the literature (e.g. ST claims that simulation is the only or at least the absolute dominant strategy). On the contrary, human social understanding rather relies on a multiplicity of strategies which are for the most part implicitly activated by contextual cues. These strategies include at least simulation strategies, theory-based inferences, and direct perception, as well as understanding based on social interaction and narratives. A plurality of social understanding was described by Andrews (2012), but she did not work out the important difference between epistemic strategies and the relevant background information which allows a systematic analysis of the rich and varying phenomena of so-called mindreading.

The second question asks how the information we obtain to understand others is stored and organized. The central claim is that information about other humans as individuals or types of persons is stored and organized in *person models*. These models are realized on two levels, namely the implicit level of person schemata and the explicit level of person images. Person models are representational structures like objects files unifying the information about an entity, e.g. another individual or a group of individuals, in a form that is less demanding than a full-fledged theory as proposed in TT.

Concisely, a model contains a unified body of information. This minimal integrated package of information enables us to understand a part of the world, e.g. by enabling us to represent an entity, such as an object, a property, a process, etc. The resulting model enables a person to represent such an entity in our world. If the information of a model is enriched over time, it unfolds into an understanding of the represented entity. In the case of understanding others, the relevant models are especially models of persons. Thus, a person model of an individual constitutes a unified body of information about the relevant individual. A person model typically has a label, namely the person's name, and it is under normal circumstances causally anchored in an entity which is ideally identical to the person in question.

[3] The notion of 'epistemic strategies' is understood in a wide sense and is here used equivalent to 'cognitive strategies', i.e. it does not imply specific high-level epistemic abilities like conscious deliberation and, thus, is not necessarily demanding. The answer to the question concerning which epistemic strategy humans use leaves still quite some room for an answer to the second question. One could in principle defend a simulation theory concerning the epistemic strategy and argue that the relevant background information for simulation is organized in person models or organized as a folk psychological theory, etc. Our view consists in the combination of the multiplicity claim concerning epistemic access and the organization of background information as person models.

(Under certain circumstances, this condition might not be fulfilled, either because something went wrong or because the respective person is non-existent, such as in the case of fictional characters). It is further argued that philosophical theories so far have tended to ignore the fact that we usually understand others by relying on rich background information concerning them and their situation. (A possible exception within the representatives of IT is Gallagher (2011).) In addition to person models we also need situation models – as we will argue below.

The two central aspects of PMT, the multiplicity of epistemic strategies and the organization of relevant background information in form of models, are explicated and motivated in more detail below.

3.2 Main Concepts

A Multiplicity of Strategies for Understanding Others

There are two main arguments employed to defend the multiplicity view concerning epistemic strategies. (i) The *ontogenetic* argument indicates that the ontogenetic development of understanding others can best be explained by describing the development of a multiplicity of epistemic strategies such that no strategy is eliminated once acquired. (ii) The *pathology* argument turns on the observation that some cases of mental disorder can best be described by demonstrating that some epistemic strategies are lacking and, thus, others – which are still available – are used as substitutions, even though they often cannot compensate for the complete lack of the original strategies.

Ad (i) The ontogenetic argument: Quite early on, babies rely on *online understanding by coordinated interaction*. They develop an expectation of an interaction scheme as demonstrated by the still-face paradigm. *Direct perception* is very relevant starting from early infancy, as proven by face-based sensitivity for and recognition of emotions based on direct perception (Zahavi 2011; Newen et al. 2015). During ontogeny, we develop further important strategies for understanding others, which also include strategies of offline understanding. It will also be indicated that we cannot observe any general dominance of one of these strategies, but that the activation of a specific strategy is dependent on the context while strategies are often activated in combination. At the age of 9 to 12 months children learn to understand others as participating in *joint attention* and *joint action* (Tomasello 1999), where the latter is demonstrated e.g. by understanding the other as following a plan like jointly constructing a Lego house (at 18 months). At 2.5 years children become sensitive to rules and norms such that they insist that group members follow rules. This involves an *understanding of others as rule-followers*, i.e. as members of the group governed by expectations concerning rule-following behaviour in relevant situations (Rakoczy et al.

2008). Furthermore, there is the well-studied ability to *understand others by explicit false beliefs* (age 4 onwards) which enables explicit theory-based inferences or explicit simulation *strategies* to understand others.[4] This is correlated with early moral understanding (see above). Finally, *understanding by explicit second-order false beliefs* develops between age 7 to 9 (Wimmer & Perner 1983). Additional epistemic strategies can be fruitfully distinguished as developing later in the process of growing up.

There is consensus that these abilities come gradually, and that abilities acquired early remain intact and in use even when more sophisticated abilities are available. To illustrate: looking at the face of a person, I may directly perceive an expression of anger. However, when I am informed that she is suffering from Parkinson's disease and therefore has severe limitations in controlling her facial expression, I will evaluate the same facial expression quite differently. Despite having a standard 'reading' of emotions from facial expressions, this new knowledge about Parkinson's disease helps me to override my spontaneous perception. Although the direct perceptual impression is still in place, I will override it in this context and use a theory-based inference to reach a new evaluation of the person, relying on other cues including the person's linguistic utterances. This also illustrates the context-dependence of the preference of one strategy over the other.

Ad (ii) The pathology argument: Some mental disorders essentially involve significant deficits in social understanding, and these cases can best be explained such that at least one strategy of the normal bundle of strategies is lacking. This can be illustrated by looking at people with Asperger's syndrome who lack an intuitive understanding of others. They are unable to directly perceive emotions based on facial expressions and they tend to avoid social interactions (Vogeley 2012). Thus, intuitive understanding by primary interaction or direct perception is (almost) unavailable for them. Since they also tend to experience themselves as being different (Vogeley 2012), they do not use simulation as a strategy. Consequently, they can only refer to theory-based inferences that might prove useful to understand others in certain situations (Kuzmanovic et al. 2011). However, they lack an intuitive generalization of this knowledge. Thus, in new or slightly modified situations, they again feel lost since they do not even have a theory on which basis to apply theory-based inferences. Since we have to deal with new or modified situations almost every day, autistic people notice their tendency to get lost and many of them avoid social encounters. This special situation is explained by the fact that in contrast to the usual availability of multiple strategies of

[4] It is an open debate how exactly theory-of-mind abilities and understanding by narratives are related to each other. While Hutto (2008) claims that the latter is more primitive than the former, we presuppose here that *understanding* by narratives is based on a theory-of-mind ability and enriches it. Thus, we do not discuss its role in addition to theory-of-mind abilities.

understanding, they are left mainly with theory-based inferences and need an explicit corpus of knowledge to apply them (since they lack intuitive generalization) (for further arguments concerning the multiplicity view, see Newen 2015a, 2015b; Fiebich & Coltheart 2015; Fiebich 2015).

In sum, social understanding usually relies on a multiplicity of epistemic strategies which are selected in a highly context-dependent manner (as demonstrated with the Parkinson case). Concerning the epistemic strategies of social understanding, we may indicate that social understanding is strongly dependent on the actual context.[5]

Person Models as Unified Information Structures (Person Files)

Having argued for the multiplicity view of epistemic strategies for social understanding, we shall focus now on the organization of the relevant background information in the form of so-called *person models*. There are only a few authors who have considered and developed an account discussed under the label '*Model Theory*' (Newen & Schlicht 2009; Maibom 2009; Godfrey-Smith 2005). The early motivation of Godfrey-Smith (2005) and Maibom (2009) offers a general answer to the status of our folk-psychological knowledge, both adopting the perspective of philosophy of science. Maibom defends a version of the claim that folk psychological knowledge has the status of a model, while understanding a model as a special version of a theory such that she remains in the camp of TT. One important advance is that she argues that a model can be based on *ordinary everyday* knowledge and need not presuppose *special* knowledge. Godfrey-Smith agrees with the latter characterization but also makes important additions by suggesting a specific understanding of 'model' which is different from Maibom's version. According to Godfrey-Smith, a model should not be understood in the tradition of a semantic view of theories. Furthermore, a model can be used in different ways such that we should distinguish between a model and its specific interpretation which he calls a 'construal'. Newen's[6] account (2015a; 2018) shares the denial of a semantic understanding of 'model' and in addition denies that a model needs to have the structure of a theory. We need a widened understanding of 'model' because this is necessary to enable us to account for the ontogeny of social understanding which is not the focus of either Maibom or Godfrey-Smith.

[5] PMT is a full-blown theory of understanding others and is has been developed in a sequence of articles (Newen/Schlicht 2009; Newen 2015a; Newen 2015b; Newen 2018). In this article, the structural organization of background knowledge in form of person models and situation models as well as their interaction is in focus.

[6] The use of the third person here indicates that the second author, Coninx, although accepting the general line of the PMT, does not accept all facets of the person model theory as presented by Newen (2015).

In line with Godfrey-Smith, Newen argues that a model can be much more parsimoniously and flexibly used, for instance by relying merely on particular parts of the model to understand certain parts of our world. The used model is not necessarily a theory of the world. Due to its ontogenetic perspective, Newen's (2015a) view on models needs to be distinguished from the claim that "one person predicts another by using a *theoretical* model" (Godfrey-Smith 2005, p. 7). Social understanding involving models can take place at different levels (see previous section) and one basic type is *online understanding by coordinated interaction*. In these cases, the model is not theoretical but perhaps just a memorized interaction schema associated with and expected in relation to a certain person. Models can be rather parsimonious information units which especially cluster information about one person (or one situation).[7] Focusing on information units about persons, their usage in early ontogeny can be without a theoretical stance: modelling a part of the world can be a different epistemic business than building a theory about it – and in early infancy, it clearly is – while models may unfold into theories during the systematic enrichment and restructuring of information.

Newen's paradigm case of a structure of a model is what is described as a mental file (Perry 1990; Recanati 2012; Newen & Marchi 2016). We can create a mental file of an object with very little information about it and start to systematically enrich and restructure the information unified in this file until it deserves to be called a concept (Newen & Marchi 2016). Since combinations of concepts constitute beliefs and combinations of beliefs constitute theories, there is a cognitive route from parsimonious models of single entities in the world to a theory about complex parts of the world. Thus, folk-psychological knowledge which is quite different and variable in structure is usually given as a model, and may unfold into a theory.

There is for the time being only one philosophical approach which aims to unfold the rather general framework of relying on 'models' into a detailed account of understanding others. This is the recent work of Newen (Newen & Schlicht 2009; Newen & Vogeley 2011; Newen 2014, 2015a, 2018). The central claim here is that relevant information about other humans as individuals or types of persons is stored and organized in person models which are either implicitly available *person schemata* or explicitly available

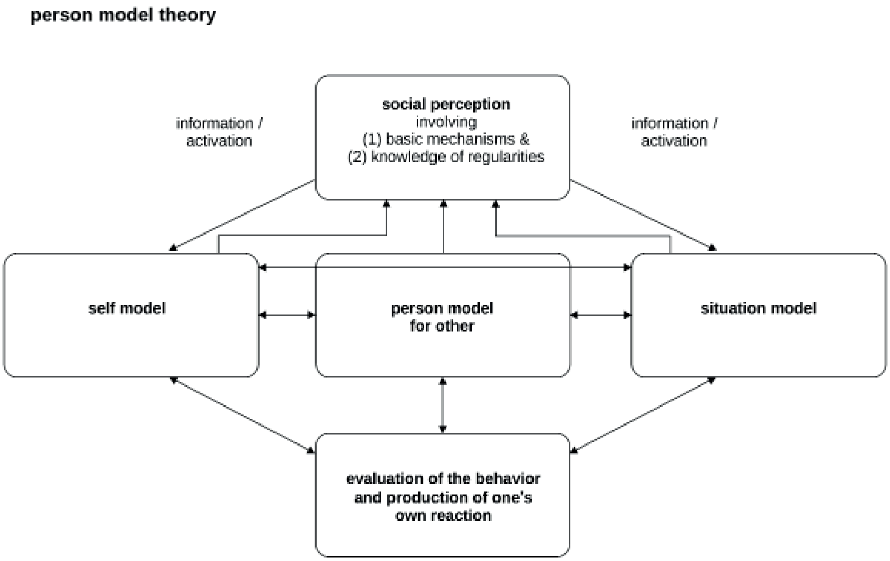
[7] Our notion of 'model' can be negatively characterized as different from semantic models which have a very constrained structure as well as from complex models in philosophy of science which always have the status of explicitly available structures. Models positively characterized are systematic informational units which integrate information about an enitity into a file which is stored in our memory system. The whole integrated information of the file or a part it can be activated (together with at least one epistemic strategy) and remain implicit to register matching properties, to trigger expectations or evaluations concerning the relevant entity: this information could (but need not) also enter an *explicit* prediction or evaluation of the relevant entity.

person images. The implicit person schema can typically be described as a unity of sensory-motor abilities and basic mental phenomena associated with one human being (or a group of humans). The schema typically functions without any explicit considerations and is activated when directly seeing or interacting with another person. In contrast to this implicit dimension, a person image is constituted by explicitly (i.e. typically consciously) available information concerning physical and mental phenomena associated with and unified to belong to one human being (or a group of humans). Thus, a person image is the unity of rather easily and explicitly available information about a person, including the person's mental setting. Both person schemata and person images can be developed for an individual, e.g. one's mother, brother, best friend etc., as well as for groups of people, e.g., anthropologists, students, medical doctors, lawyers, etc. Furthermore, person models are not only created for other people, but also for oneself.

There is recent empirical evidence from neuroscience that we actually construct and rely on person models (Hassabis et al. 2013, see Newen 2015a, section 5.3). It has been shown that there are neural correlates of imagining two central features of the 'big five' in personality psychology, i.e. 'agreeable' in contrast to anti-social personalities, and 'introvert' in contrast to extrovert personalities. Furthermore, it was shown that the combinations of personality types like 'agreeable-ness' and 'extroversion' are represented in a systematic modulation of the medial prefrontal cortex.

Situation models and their intertwinement with person models

An account of full-blown PMT must mention one further component, namely situation models. Humans have the ability to understand others by completely abstracting from the individual: e.g. it can be sufficient to predict the behaviour of a restaurant guest that we expect her to act according to the conventions of a high-class restaurant. This type of understanding is developed together with rule-based understanding of others at the age of 2.5 years (see above). In new contexts, especially in new cultural contexts, we begin by employing an understanding of others mainly on the basis of noticing rule-based behaviour which we discern as being adequate in a situation. For instance, as a European one learns to understand other restaurant guests by learning the rule-based behaviour characteristic of high-class restaurant situations in Japan (special greetings, taking off shoes, sitting in a special way, etc.). Thus, we not only create person models but also situation models, and our understanding of others uses both types of model as input and selects the model most helpful for evaluating the other person's behaviour. In the following, we explicate in two steps, first, why situation models are necessary and, second, how person models and situation models interact in the social evaluation process (figure adapted from Newen 2015a, p. 21).



The relevance of situation models is based on the observation that situation models are sufficient to understand others, if we need not account for the individual person we aim to understand. If one is not interested in another as an individual, but merely as another agent in a situation, the situation itself often offers sufficient information to predict the behaviour of the people and to coordinate one's own action with theirs, e.g. many shopping interactions are of this type. Furthermore, situation models allow us to predict the behaviour of all people fulfilling typical roles in the situation, e.g. the role of the guest, the seller or the cleaning person within a restaurant. Since humans frequently need to coordinate actions with many persons in the same situation, understanding others on the basis of situation models is a very important tool for life in larger groups. The distinction between person models and situation models is also captured in the difference between reason explanations which focus on consideration of an individual, on the one hand, and causal history explanations which highlight the relevant situation and how it developed, on the other hand (Malle et al. 2006, Fiebich 2015).

Thus, a full-fledged theory of understanding others needs to include situation models as well as basic ideas of the interdependence of personal models and situation models. A situation model is a pattern constituted by a sequence of typical activities or events in a type of situation involving human agents whereby the agents are only represented in an unspecific way comparable to variables in logic. Paradigmatically, this includes situations, such as entering a fast-food restaurant to arrange your lunch, entering a class room to participate in a university seminar, entering a bar to meet friends, etc. We need situation models to coordinate quickly with others according to social expectations in such situations, even if we do not know the persons involved. Of course, we need to account for the fact that situation models vary intensely across cultures, e.g. how you have to behave in a restaurant is quite different in Japan and the United States.

Finally, we will briefly illustrate the second aspect, i.e. how person models and situation models interact. If we consider once more the restaurant situation, then a typical case could be that while waiting in the queue to order lunch, one starts to communicate with the person in front. This immediately initiates a basic person perception (Macrae & Quadflieg 2010; Newen 2015a) which leads to the creation of a person model, at least, in the working memory. Whether it is transferred into long-term memory depends on attentional features that rely on the estimated relevance of this interaction for future life. For instance, if the person in front impresses us a lot or if we discover that she will start working in the same company, this would lead to the creation of an explicit person model, i.e. a person image, which is enriched step by step during each encounter. If we have a minimally rich person model, it is cognitively economical to rely on this person model whenever interacting with the person. It allows for much better predictions because we can account not only for the general information concerning a situation type, but also for the more fine-grained information about this very person.

Situation models can be enriched by person models and the other way around: a situation model is a pattern constituted by a sequence of typical activities or events in a type of situation involving human agents whereby the agents are only represented in an unspecific way comparable to variables in logic. Person models can enrich these variables in a way that can be compared to substituting a variable by a logical constant (by fitting the person representations into the unspecific agent slots of the event structure). If one engages in a social understanding with an activation of a person model, it can of course be naturally enriched by integrating the person model into a relevant situation model (by including person representations into an event sequence). The richest understanding of a social situation is possible, if we have a detailed situation model. For instance, when we are at a formal birthday party, and we know each member of the party, we can enrich the situation model with all relevant person models. This allows us to deliver detailed explanations and make detailed predictions. To sum up: We rely on both person models and situation models. Situations models are more important for basic social perceptions of situation including the agents involved in the situation (but ignoring them as individuals), while person models are especially fruitful, if an understanding of the individuals themselves is relevant. The richest understanding demands a combination of both and their integration in a structurally fitting manner.

3.3 General Profile and Advantages of PMT

The person model theory (PMT) contains both, a theory about the epistemic strategies involved in social understanding and a theory about the organizational structure of relevant background information, either in person models or situation models. The resulting general picture of social understanding is the following. Social understanding needs the activation of at least one epistemic strategy. Epistemic strategies can be used individually or in combination as they gradually occur in ontogeny. Epistemic strategies are activated in a particular situation by social cues. This activation can be based on a single social cue, e.g. biological motion, but typically, it relies on the perception of many social cues at once, e.g. in the case of emotion recognition based on facial expression, body posture, gestures. Moreover, in the activation of a particular epistemic strategy typically person models and/or situation models are involved. Social understanding does not merely rely on directly perceivable social cues but also on background information organized in models. Person models and situation models of certain entities also unfold ontogenetically (Newen & Marchi 2016) and are systematically enriched by the new information a person receives. Thus, these models are not rigid but dynamically developing.[8] This enables flexible and reliable social registration, prediction and evaluation. The interaction of epistemic strategies, person models and situation models enables thereby a great variety of types of social understanding. In a simplified overview, we distinguished ontogenetically three main types of social understanding:

(i) *online understanding of others* realized as coordinated interaction or as participating in joint attention and joint action, typically based on intuitive epistemic strategies like direct perception or low-level simulation in combination with person model information, e.g. a child smoothly interacting with its mother activating an interaction schema or joint attention concerning an object.

(ii) *understanding of others as rule-followers* realized as expecting others to follow rules which are constitutive for members of a social group in a specific situation. This is based on person models for types of individuals which are also called person models of groups (Newen 2015a), e.g. when we recognize a person as a member of a soccer club, we expect him to be a good soccer player preparing for a game when it is time to do so or to help preparing a club party when the celebration of the local championship is announced. These examples illustrate that person models of groups (types of individuals) are essentially involved and they have to be combined with situation models to activate specific expectations. Furthermore, some epistemic strategy has to be activated and this can be any of the available strategies. Such a flexible use of the most efficient epistemic strategy is also presupposed for the third type of social understanding:

[8] This implies that the epistemic strategies are not only used to register and evaluate a social situation but they are at the same time a tool of adjusting the contents of person models and situation models.

(iii) *understanding of others as having an individual mindset of attitudes* realized by the attribution of beliefs, desires, hopes, fears, etc. to others in a certain situation. This is based on person models of individuals in a specific situation. Thus, persons are represented as having individual beliefs and desires such that e.g. they can be predicted not to behave according to a relevant social roles they have in a situation but according to their individual mindset which can differ from relevant social expectations.

In the scope of this article, the PMT cannot be outlined in more detail. However, it has been illustrated that PMT clearly differs from its competitors. PMT can account for the plurality and development of epistemic strategies employed by a person in different situations with regard to different agents and in different stages of ontogenetic development. In addition, the introduction of the person model enables the understanding of several important aspects which at least one of the competitors fails to account for:

1. The person model theory can convincingly account for the difference between understanding a complete stranger by relying on a situation model and understanding a well-known familiar person by relying on a rich and more specific person model. No other theory can account for the systematic understanding of individual idiosyncrasies and the relevance of cultural schemata of how to behaviour in a particular context. On the contrary, person models and situation models can do the job.
2. By appealing to the distinction between implicit and explicit person models, PMT can account for the difference between basic or intuitive understanding and complex or theory-based understanding of others which is underdeveloped in TT.
3. With the difference between a person model of oneself and person models of others, PMT can account for an understanding of others which goes beyond the own-self model as the sole source of understanding others, contrary to ST. PMT can account both for an understanding of others based on the self-model and an understanding of others based on the person model of other individuals or types of individuals which can be radically different from the self-model.
4. PMT differs from IT, since it addresses not only basic online understanding, but also offline social understanding.
5. Furthermore, with the outlined dynamics between situation models and person models, PMT especially offers a tool to account for situational and personal features as well as the cultural variation of their relevance. Thus, it seems correct to call PMT a new approach, not just a variant of an existing one.

3.4 Challenges

PMT is a new framework to account for understanding others which confers several advantages compared to the alternative accounts. Nevertheless, it is still accompanied by open questions and challenges. This includes among others a more precise investigation of how the different epistemic strategies interact and under which conditions one strategy is preferred in case it conflicts with another. Moreover, there is a need to clarify further how person models are individuated and how they are cognitively implemented. While in the published work, there is a description of an fMRI study of Hassabis et al. (2013) that provides evidence for person models as models of personality types (person schemata for groups), it would be helpful to provide similar evidence concerning the implementation of person models of individuals. PMT also needs an explication of the borderline between person models of groups which are already general and rather general knowledge of folk psychological rules as described by TT. Finally, a detailed description of the interaction between situation models and person models is needed as they strongly influence each other: a person often has different dispositions to behave, to the point of virtually different personality traits showing up depending on the situation. In a job situation a person may be extremely harsh, while being friendly in a family context.

4. Concluding remarks

We discussed the four main theories of understanding others: ST, TT, IT and PMT. While the first three accounts – which have indeed been under discussion in the philosophical literature for quite some time – reveal critical gaps, PMT offers a promising attempt to close these gaps, albeit still having open questions that its defenders have to answer. The future of the debate about social cognition will tell which theory is the most fruitful framework and how it should be unfolded to deliver the most adequate descriptions and predictions. This is clearly an interdisciplinary challenge which requires the combination of insights at least from philosophy, psychology, psychiatry and neuroscience. The current state of knowledge indicates the need for a multidimensional and flexible understanding of human mindreading which involves individuals, groups, cultures and situations.

References

Adolphs, R., D. Tranel, H. Damasio & A. Damasio 1994. “Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala.” *Nature* 372 (6507): 669–72.

Andrews, K. 2012. *Do apes read minds? Towards a new folk psychology*. Cambridge (MA): MIT Press.

Apperly, I. & S. Butterfill 2009. "Do humans have two systems to track beliefs and belief-like states?" *Psychological Review* 116 (4): 953–970.

Baron-Cohen, S. 1995. *Mindblindness – An Essay on Autism and Theory of Mind*. Cambridge (MA); London: The MIT Press.

Becchio, C., M. Adenzato & B. G. Bara 2006. "How the brain understands intention: different neural circuits identify the componential features of motor and prior intentions." *Consciousness and Cognition* 15 (1): 64–74.

Bertin, E. & T. Striano 2006. "The still-face response in newborn. 1.5-, and 3-month-old infants." *Infant Behavior and Development* 29 (2): 294–297.

Birch, S.A. & P. Bloom 2003. "Children are Cursed: An Asymmetric Bias in Mental-State Attribution." *Psychological Science* 14 (3): 283–6.

Bohl, V. 2015. "Continuing debates on direct social perception: Some notes on Gallagher's analysis of 'the new hybrids.'" *Consciousness and Cognition* 36: 466–71.

Butterfill, S. A. 2009. "Seeing causings and hearing gestures." *Philosophical Quaterly* 59 (236): 405–28.

Camerer, C., G. Loewenstein & M. Weber 1989. "The Curse of Knowledge in Economic Settings: An Experimental Analysis." *Journal of Political Economy* 97 (5): 1232–54.

Carruthers, P. 2015. "Perceiving Mental States." *Consciousness and Cognition* 36: 498–507.

Davies, M. & T. Stone 1995. *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers.

De Bruin, L. & A. Newen 2012. "An Association Account of False Belief Understanding." *Cognition* 123 (2): 240–259.

De Bruin, L., M. van Elk & A. Newen 2012. "Reconceptualizing Second-Person Interaction." *Frontiers in Neuroscience* 6: 151.

De Jaegher, H. and Di Paolo, E. 2007. "Participatory sense-making: An enactive approach to social cognition." *Phenomenology and the Cognitive Sciences* 6 (4): 485–507.

Di Paolo, E. & H. De Jaegher 2012. "The interactive brain hypothesis." *Frontiers in Human Neuroscience* 6: 1–16.

Di Pellegrino, G., L. Fadiga, L. Fogassi, V. Gallese & G. Rizzolatti 1992. "Understanding motor events: a neurophysiological study." *Experimental Brain Research* 91 (1): 176–80. Retrieved from http://link.springer.com/article/10.1007/BF00230027

Ekman, P. & W. V. Friesen 1971. "Constants across culture in the face and emotion." *Journal of Personality and Social Psychology* 17: 124–129.

Fiebich, A. 2015. *Varieties of Social Understanding*. Münster: mentis Verlag.

Fiebich, A. & Coltheart, M. 2015. "Various ways to understand other minds. Towards a pluralistic approach to the explanation of social understanding." *Mind and Language* 30 (3): 235–58.

Fisher, J. C. 2006. "Does Simulation Theory Really Involve Simulation?" *Philosophical Psychology* 19 (4): 417–32.

Frith, U., & Frith, C. D. 2003. "Development and neurophysiology of mentalizing." *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 358 (1431): 459–473.

Furley, P., T. Moll & D. Memmert 2015. "'Put your Hands up in the Air'? The interpersonal effects of pride and shame expressions on opponents and teammates." *Frontiers in psychology* 6: 1361.

Gallagher, S. 2001. "The practice of mind: Theory, simulation, or interaction?" *Journal of Consciousness Studies* 8 (5–7): 83–107.

Gallagher, S. 2007. "Simulation Trouble." *Social Neuroscience* 2 (3&4): 653–65.

Gallagher, S. 2008. "Direct perception in the intersubjective context." *Consciousness and Cognition* 17 (2): 535–43.

Gallese, V. & A. Goldman 1998. "Mirror neurons and the simulation theory of mind-reading." *Trends in Cognitive Sciences* 2 (12): 493–501. Retrieved from http://www.sciencedirect.com/ science/article/pii/S1364661398012625

Gao, T., Newman, G. E., and Scholl, B. J. 2009. "The psychophysics of chasing: A case study in the perception of animacy." *Cognitive Psychology* 59 (2): 154–79.

Godfrey-Smith, P. 2005. "Folk psychology as a model." *Philosophers' Imprint*, 5 (6): 1–16.

Goldman, A. I. 1989. "Interpretation Psychologized." *Mind & Language* 4 (3): 161–85.

Goldman, A. I. 1992. "In Defense of the Simulation Theory." *Mind & Language* 7 (1&2): 104–19.

Goldman, A. I. 2006. *Simulating Minds – The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goldman, A. I. & L. C. Jordan 2013. „Mindreading by simulation: The roles of imagination and mirroring." In *Understanding Other Minds*, 3rd edition, edited by S. Baron-Cohen, M. Lombardo & H. Tager-Flusberg, 448–66. Oxford, Oxford University Press.

Goldman, A. I. & N. Sebanz 2005. "Simulation, mirroring, and a different argument from error." *Trends in Cognitive Sciences* 9 (7): 320; author reply 321.

Gopnik, A. & H. M. Wellman 1992. "Why the Child's Theory of Mind Really Is a Theory." *Mind & Language* 7 (1&2): 145–71.

Gordon, R. M. 1986. "Folk Psychology as Simulation." *Mind & Language* 1 (2): 158–71.

Gordon, R. M. 1992. "The Simulation Theory: Objections and Misconceptions." *Mind & Language* 7 (1&2): 11–34.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., and Schacter, D. L. 2013. "Imagine all the people: how the brain creates and uses personality models to predict behavior." *Cerebral Cortex*, March 5, 2013. doi:10.1093/cercor/bht042.

Herschbach, M. 2012. "On the role of social action in social cognition: a mechanistic alternative to enactivism." *Phenomenology and the Cognitive Sciences* 11 (4): 467–486.

Hutto, D. D. 2007. "The narrative practice hypothesis: origins and applications of folk psychology." *Royal Institute of Philosophy Supplement* 60: 43–68.

Hutto, D. 2008. *Folk-psychological narratives.* Cambridge (MA): MIT Press.

Hutto, D. & S. Gallagher 2008. "Understanding others through primary interaction and narrative practice." In *The Shared Mind: Perspectives on Intersubjectivity*, Converging Evidence in Language and Communication Research 12, edited by J. Zlatey, T. Racine, C. Sinha & E. Itkonen, 17–38. Amsterdam: John Benjamins Publishing Company.

Keysers, C. & V. Gazzola 2009. "Unifying Social Cognition." *Progress in Brain Research* 156: 379–401.

Kilner, J. M., K. J. Friston & C. D. Frith 2007. "The mirror-neuron system: a Bayesian perspective." *Neuroreport* 18 (6): 619–23.

Kuzmanovic, B., Schlibach, L., Lehnhardt, F., and Vogeley, K. 2011. "A matter of words: Impact of verbal and nonverbal information on impression formation in high-functioning autism." *Research in Autism Spectrum Disorders* 5 (1): 604–13.

Macrae, C. N. & Bodenhausen, G. V. 2000. Social cognition: Thinking categorically about others. *Annual Review of Psychology*, 51, 93–120.

Macrae, C. N. & Quadflieg, S. 2010. Perceiving people. In S. Fiske, D. T. Gilbert & G. Lindzey (Eds.) *Handbook of Social Psychology*. New York, NY: McGraw-Hill, 428–463.

Malle, B. F. 2006. "The actor-observer asymmetry in attribution: a (surprising) meta-analysis." *Psychological Bulletin*, 132 (6), 895–919.

Marchi, F. & A. Newen 2015. "Cognitive penetrability and emotion recognition in human facial expressions." *Frontiers in Psychology* 6 (828). doi: 10.3389/fpsyg.2015.00828

MacPherson, F. 2012. “Cognitive penetration of colour experience. Rethinking the issue in light of an indirect mechanism.” *Philosophy and Phenomenological Research* 84 (1): 24–62.

Maibom, H. L. 2009. “In defence of (model) theory theory.” *Journal of Consciousness Studies* 16 (6–8): 360–78.

Newen, A. 2014. “Selbst– und Fremdverstehen: die Personenmodelltheorie als Analyserahmen für mentale Stӧrungen.“ In *Verleumdung und Verrat: Dissoziative Störungen bei schwer traumatisierten Menschen in Folge von Vertrauensbrüchen,* edited by Vogt, R., Krӧning: Asanger, 209–218.

Newen, A. 2015a. “Understanding Others. The Person Model Theory” In *Open MIND*, 26, edited by J. M. Metzinger & T. Windt, 1–28. Frankfurt am Main: MIND Group. Open.mind.net/ doi: 10.15502/9783958570320, 1–28 [Printed version: Newen, A. (2016): “Understanding Others – The Person Model Theory.” In: Metzinger, T. & Windt, J.M. (eds.): *Open MIND. Philosophy and the Mind Sciences in the 21st Century,* Cambridge MIT Press, S. 1049–1076].

Newen, A. 2015b. “A Multiplicity View for Social Cognition: Defending a Coherent Framework.” In: *Open MIND*, 26, edited by J. M. Metzinger & T. Windt. Frankfurt am Main: MIND Group. http://open-mind.net/ doi: 10.15502/9783958570320 [Printed Version: Newen, A. (2016): A Multiplicity View for Social Cognition: Defending a Coherent Framework – A reply to Lisa Quandt. In: Metzinger, T. & Windt, J.M. (Hrsg.): *Open MIND. Philosophy and the Mind Sciences in the 21st Century,* Cambridge MIT Press, 1095–1102].

Newen, A. 2017. “Defending the liberal-content view of perceptual experience: Direct social perception of emotions and person impressions.” *Synthese* 194 (3): 761–785. doi: 10.1007/s11229–016–1030–3.

Newen, A, 2018. “The Person Model Theory and The Question of Situatedness of Social Understanding. In: Newen, A., de Bruin, L., Gallagher, S. (eds): *Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press, 469–492.

Newen, A. & F. Marchi 2016. “Concepts and their organizational structure.” In *Concepts and Categorization,* edited by D. Hommen, C. Kann and T. Osswald, Münster: mentis Verlag, 197–227.

Newen, A. & T. Schlicht 2009. “Understanding Other Minds: A criticism of Goldman’s simulation theory and an outline of the Person Model Theory.” *Grazer Philosophische Studien* 79: 209–42.

Newen, A. & P. Vetter 2016. „Why cognitive penetration of our perceptual experience is still the most plausible account.” *Consciousness and Cognition* 47: 26–37.

Newen, A. and Vogeley, K. 2011. “Den anderen verstehen.“ *Spektrum der Wissenschaft* 8/2011.

Newen, A., A. Welpinghus & G. Juckel 2015. “Emotion Recognition as Pattern Recognition: The Relevance of Perception.” *Mind & Language* 30 (2): 187–208.

Neufeld, E., Brown, E. C., Lee-Grimm, S.-I., Newen, A., Brüne, M. (2016). Intentional Processing Results from Automatic Bottom-Up Attention: An EEG-Investigation into the Social Relevance Hypothesis Using Hypnosis. *Consciousness and Cognition* 42, 101–112. doi: 10.1016/j.concog.2016.03.002

Nickerson, R. S. 1999. “How we know – and sometimes misjudge – What others know: Imputing one's own knowledge to others.” *Psychological Bulletin* 125 (6): 737–59.

Onishi, K. H. & R. Baillargeon 2005. “Do 15-month-old infants understand false beliefs?” *Science* 308 (8), 255–258.

Pacherie, E. 2005. “Perceiving intentions.” In *A explicação da interpretação humana,* edited by J. Sàágua, Lisbon: Edições Colibri, pp.401–14.

Perry, J. 1990. “Self-Notions.” *Logos* 11: 17–31.

Rakoczy, H., Warneken, F., Tomasello, M. 2008. “The Sources of Normatively: Young children's awareness of the normative structure of games.” *Developmental Psychology* 44 (3): 875–881.

Recanati, F. 2012. *Mental Files.* Oxford, UK: Oxford University Press.

Rizzolatti, G. & L. Craighero 2004. “The mirror-neuron system.” *Annual Review of Neurosc.* 27: 169–92.

Rutherford, M. D. and Kuhlmeier, V. A. 2013. *Social perception: Detection and interpretation of animacy, agency, and intention.* Cambridge (MA): MIT Press.

Saxe, R. 2005. “Against simulation: the argument from error.” *Trends in Cognitive Sciences* 9 (4): 174–9.

Scheler, M. 1954. *The Nature of Sympathy.* New Haven: Yale University Press.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., Vogeley, Kai. 2013. “Toward a second-person neuroscience.” *Behavioral and Brain Sciences* 36: 393–462. doi:10.1017/S0140525X12000660

Siegel, S. 2009. “The visual experience of causation”. *Philosophical Quaterly* 59 (236): 519–40.

Sinigaglia, C., G. Rizzolatti 2008. *Mirrors in the Brain: How Our Mind Share Actions and Emotions.* Oxford: Oxford University Press.

Spaulding, S. 2010. “Simulation theory.” *Wiley Interdisciplinary Reviews: Cognitive Science* 1 (4): 527–38.

Sodian, B., C. Thoermer & U. Metz 2007. "Now I see it but you don't: 14-month-olds can represent another person's visual perspective." *Developmental Science* 10 (2): 199–204.

Sorce, J., R. Emde, J. Campos & M. Klinnert 1985. "Maternal Emotional Signaling: Its Effect on the Visual Cliff Behavior of 1-Year-Olds." *Developmental Psychology* 21 (1): 195–200.

Stich, S. & S. Nichols 1995. "Second Thoughts on Simulation." In *Mental Simulation: Evaluations and Applications*, edited by A. Stone & M. Davies, 87–108. Oxford: Blackwell.

Tomasello, M. 1995. "Joint attention as social cognition." In *Joint attention: Its origins and role in development*, edited by C. Moore & P. J. Dunham, 103–130. Hillsdale: Lawrence Erlbaum.

Tomasello, M. 1999. *The cultural origins of human cognition.* Cambridge (MA): Harvard University Press.

Trevarthen, C. B. 1979. "Communication and cooperation in early infancy: A description of primary intersubjectivity." In *Before speech: The Beginning of Interpersonal Communication*, edited by M. Bullowa, 321–48. Cambridge: Cambridge University Press.

Trevarthen, C. B. 2006. "The concepts and foundations of intersubjectivity." In *Intersubjective Communication and Emotion in Early Ontogeny,* edited by S. Braten, 15–46. Cambridge: Cambridge University Press.

Trevarthen, C. B. & P. Hubley 1978. "Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year." In *Action, gesture and symbol: The emergence of language*, edited by A. Lock, 183–229. London: Academic.

Tversky, A. & D. Kahneman 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31.

Tye, M. 1995. *Ten problems of consciousness. A representational theory of the phenomenal mind.* Cambridge, MA: MIT Press.

Vetter, P. & Newen, A. 2014. "Varieties of cognitive penetration in visual perception." *Consciousness & Cognition* 27: 62–75. Doi:10.1111/j.1468–0017.2006.00298.x.

Vogeley, K. 2012. *Anders sein – Hochfunktionaler Autismus im Erwachsenenalter*. Weinheim: Beltz-Verlag.

Vogeley, K., P. Bussfeld, A. Newen, S. Herrmann, F. Happé, P. Falkai, W. Maier, N. J. Shah, G. R. Fink & K. Zilles 2001. "Mind reading: neural mechanisms of theory of mind and self-perspective." *Neuroimage* 14 (1 PT 1): 170–81.

Vogeley, K. & A. Newen 2002. "Mirror neurons and the self construct." In *Mirror Neurons and the Evolution of Brain and Language,* edited by Maxim

I. Stavenov and Vittorio Gallese, 135–150. Amsterdam: John Benjamins Publishing.

Vogeley, K., L. Schilbach & A. Newen 2013. "Soziale Kognition." *Interdisziplinäre Anthropologie* 1: 13–40.

Weinberg, M. K., M. Beeghly, K. L. Olson & E. Tronick 2008. "A Still-face Paradigm for Young Children: 2½ Year-olds' Reactions to Maternal Unavailability during the Still-face." *Journal of Developmental Processes* 3 (1): 4–22.

Wicker, B., C. Keyers, J. Plailly, J. P. Royet, V. Gallese & G. Rizolatti 2003. "Both of us disgusted in *My* insula: the common neural basis of seeing and feeling disgust." *Neuron* 40 (3): 655–64.

Wimmer, H. & J. Perner 1983. "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception." *Cognition* 13: 103–28.

Zahavi, D. 2011. "Empathy and direct social perception: A phenomenological proposal." *Review of Philosophy and Psychology* 2 (3): 541–58.

Mario De Caro
Università Roma Tre & Tufts University

Original Scientific Paper
UDK 165 Патнам Х.
14 Патнам Х.
165.7/.8

PUTNAM ON THE MIND-BODY PROBLEM

Abstract. *This article discusses Hilary Putnam's views on the mind-body problem, by locating them in the general context of a satisfying pluralistic naturalism that he tried to articulate throughout his entire philosophical career. The first attempt in this direction was* computational functionalism, *his version of psychological functionalism centered on the analogy between mind/body and software/hardware, which (differently from David Lewis and others) he came to think of as an empirical hypothesis. That was a very successful proposal; however, later Putnam abandoned it and embraced what he called "liberal functionalism". The reason for this change of mind was twofold: on the one hand, Putnam reached the conclusion that computational functionalism was incompatible with his views on semantic externalism; on the other hand, he began to think that mental states, besides being compositionally plastic (i.e., two entities can be in the same psychological state without being in the same physical state), are also computationally plastic (i.e., two entities can be in the same psychological state without being in the same functional state). In conclusion, I will argue that "liberal functionalism" opened an interesting perspective for a successful non-reductive version of naturalism.*

Key words: *mind-body problem; Hillary Putnam; functionalism; plasticity*

1. From computational functionalism to liberal functionalism

Hilary Putnam was an omnivorous philosopher: paraphrasing Terence's famous words, one could say that nothing philosophical was alien to him. From philosophy of mind to epistemology, logic to philosophy of language, history of philosophy to metaphysics, philosophy of religion to ethics, he offered contributions that were always brilliant and often seminal. However, as is well known, Putnam's huge philosophical work was marked by frequent changes of mind; therefore, one may think that his philosophical development was devoid of continuity. This interpretation, however, would be mistaken, since several well-defined major threads and unchanged goals unified his entire philosophical production – and this is particularly true of his reflection on the mind-body problem, which is the subject of this paper. Therefore, in order to understand the trajectory of Putnam's thought regarding the mind-body problem, one has to consider three other main themes which remained constant in most of his work:

1. The search (which began at the beginning of the Sixties) for an encompassing form of *liberal naturalism*, able to do justice to the scientific worldview, on the one hand, and the manifest worldview with his ineliminable normative components, on the other hand;

doi: 10.5937/BPA1831155C

2. The pursuit of a *scientific realism* immune from the insurmountable difficulties of the so-called "*metaphysical realism*" (the idea that there is exactly one true and complete description of the way the world is);
3. An uninterrupted allegiance to *semantic externalist* (which started at the beginning of the Seventies);

1.1. Computational functionalism

Computational functionalism was Putnam's first accomplished attempt to answer the mind/body problem.[1] The target that Putnam had in mind when he developed his functionalist view (or "computational functionalism", as he came to call it later) was the so-called "mind-brain type identity view" (proposed by Smart, Place, and Feigl), according to which mental types of events are identical to physical (that is, cerebral) types of events. Putnam offered a famous argument against that view. Take a human's mental event, such as feeling pain. Now, imagine an octopus having pain. The octopus is in the same mental state of the human, but certainly its physical state is very different from the human's physical state (in fact, the octopus has a brain that is anatomically very different from ours).[2] Moreover, besides octopuses, also a huge number of animals that feel pain have brains and nervous systems very different from ours and from each other (and nowadays we even have some evidence that even vegetables may feel pain).[3] But there is more: we can imagine that also robots or aliens physically very different from us could be able to feel pain – and it may well be that these entities are possible. So, if it is true that even a robot, made of silicon instead of carbon, or an alien, made of who knows what, could feel pain, it means that there is a potentially unlimited number of physical bases of pain. Consequently, there cannot be a type-identity between the mental and the physical, because the physical type in question is unavoidably open-ended.

An alternative view, which Putnam started to elaborate in 1960, was based on the view that the mind should be interpreted as a Turing machine or as a piece of software running on some hardware (the brain). The mind, in this sense, is a program hardwired in a physical basis (for humans, the brain) and it defines all the mental states, which are seen as intrinsically computational; and, more specifically, each mental state is functionally defined by its causal inputs and outputs, independently of its physical base. In this light, if a mental

[1] See the last seven essays in H. Putnam, *Mind, Language, and Reality. Philosophical Papers*, vol. II, Cambridge University Press, Cambridge 1975.

[2] "In an octopus, it is not clear where the brain itself begins and ends. The octopus is suffused with nervousness" (P. Godfrey-Smith, "The Mind of an Octopus", *Scientific American Mind*, 28, 1, 2017, pp. 62–69; retrieved at https://www.scientificamerican.com/article/the-mind-of-an-octopus/).

[3] Frank Kühnemann of the Institute for Applied Physics in Bonn has found out that "when a leaf or a stem is cut off, the plant 'cries out' in pain by releasing the gas ethylene over its entire surface": see https://www.dw.com/en/when-plants-say-ouch/a-510552–1.

state of a human and one of a robot play the same causal role, the human and the robot are in the same computational (i.e., mental) state.

1.2. Liberal functionalism

At the beginning, Putnam interpreted computational functionalism as valid a priori (as long as there minds exist, of course), but later he came to see it as a (strongly corroborated) empirical hypothesis. However, finally, he saw insurmountable difficulties with computational functionalism as such and abandoned it altogether. This is how Putnam summarized how his attitude toward this view changed:

> In "Minds and Machines" I assumed that the brains of both robots (pretend there are intelligent ones!) and humans can be described as computers. I suggested, but didn't commit myself to, the idea that the mental states of those robots and those humans could be identified with what I called the "logical states" of their brains, meaning by that the states described by their programs. I called them "logical states" to emphasize that the *physical description* of those states was irrelevant; if a robot "brain" and a human brain have the same program, then the human and the robot have the same mental states, if that idea is right. Subsequently, I committed myself to this identification of mental states and computational states as an "empirical hypothesis". Very soon, I found difficulties with the identification of mental states and computational states—difficulties that led me to various reformulations.[4] Eventually, I found I couldn't reconcile this identification with my advocacy of externalist and anti-individualist semantics in "The Meaning of 'Meaning'", and I finally I discarded it as "science fiction".[5]

As Putnam hints in this passage, the main reason for which he abandoned computational functionalism was that he realized its incompatibility with another view he had been defending since the beginning of the Seventies: semantic externalism. Putnam presented the latter view by appealing to the famous "Twin Earth" thought-experiment, which started the "externalist revolution" (to paraphrase John Heil 1992, 24). Here is how Putnam summarizes that thought experiment:

> Imagine a planet like earth—call it "Twin Earth"—on which the liquid that fills the lakes and rivers, that people drink, etc. is not H_2O but a different compound XYZ, with similar superficial characteristics. The Twin Earthers are supposed to be our "Doppelgangers"; some of

[4] Putnam describes these reformulations in his autobiographical essay in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Blackwell, Oxford 1993, pp. 507–513.

[5] H. Putnam, "Corresponding to Reality", in H. Putnam (eds.), *Philosophy in an Age of Science*, ed. by M. De Caro and D. Macarthur, Harvard University Press, Cambridge (MA) 2012, pp. 72–90; quotation at pp. 39–40.

> them even speak English. Also, I imagine the year to be 1750, hence prior to the chemical composition of either water or twater (Twin Earth 'water') being known. The English speaking Twin Earthers naturally call twater "water" (and the French-speaking ones call it "eau", and the German-speaking ones call it "Wasser"). The linguistic intuition of the great majority of people who have considered this thought experiment is that upon learning that Twin Earth "water" doesn't consist of H_2O at all, we Earthers would say "it isn't really water". The word "water" has a different meaning on Earth and on Twin Earth. Twin Earthian Oscar's word "water" and Earthian Oscar's word "water" are homonyms, but not synonyms. They do not have the same meaning—not even if Earthian Oscar and Twin Earthian Oscar happen to be microphysical duplicates![6]

The famous slogan that Putnam derived from this view was "Meanings ain't in the head!"; however, the slogan was too prudent, as Putnam repeatedly said later, since it should have rather be "Thoughts ain't in the head!". The externalist view that was supported by the Twin Earth experiment concerned our causal interactions with the physical world: so it could be called "physical externalism". However, Putnam also defended another version of externalism, "social externalism", on the basis of the idea of the "linguistic division of labor", according to which the necessary and sufficient conditions for individuating the referents of a general name (such as "elm" or "aluminum") are "all present in the linguistic community considered as a collective body; but that collective body divides the 'labor' of knowing and employing these various parts of the 'meaning' of [that general name]"[7].

The reason why semantic externalism is incompatible with computational functionalism is not difficult to understand, even if Putnam resisted several years before drawing such conclusion. As said, according to computational functionalism, thoughts are internal to the mind on the individuals; according to semantic externalism, instead, thoughts reach out the individual minds, since they involve our transactions with the external world (natural and social). Writes Putnam:

> I had to give up "functionalism," for example, that is, the doctrine that our mental states are just our *computational* states (as implicitly defined by a "program" that our brains are hard-wired to "run"), because that view is incompatible with the semantic externalism that years of thinking about the topic of reference had eventually led me to develop. If, as I said in "The Meaning of 'Meaning,'" our intentional mental states aren't in our heads, but are rather to be thought of *as world-involving abilities*, abilities identified by the sorts of transactions

[6] H. Putnam, "The Development of Externalist Semantics", in H. Putnam, *Naturalism, Realism, and Normativity*, ed. by M. De Caro, Harvard University Press, Cambridge (MA) 2016, pp. 199–212.

[7] H. Putnam, "The Meaning of Meaning", in Id., *Mind, Language and Reality,* cit., pp. 215–271; quotation at pp. 227–228.

> with our environment that they facilitate, then they aren't identified simply by the "software" of the brain.

Putnam justified his abandonment of computational functionalism also in another way. In fact, he came to realize that mental states "are not only compositionally plastic, that is, capable in principle of being realized in different sorts of hardware, but *computationally plastic*, that is, capable of being realized in different sorts of software."[8] Mental attitudes such as "believing (or fearing or hoping) something" can be mapped onto, or realized by, a software in many different ways.[9]

However, Putnam did not entirely abandon functionalism: he rather reinterpreted it in an externalistic spirit, and called the resulting new view "liberal functionalism". The idea behind the new view was that the mind is a system of object-involving abilities, which still are functions – but not functions that are merely hardwired in the brain of a speaker (as it was for computational functionalism), but functions that are intrinsically "transactional", since from the start they involve the natural and social environment in which that speaker is located.

> I still believe that our so-called "mental states" are best thought of as *capacities to function*, but not in the strongly reductionist sense that went with the model of those states as "the brain's software." They are, so to speak, "long-armed" functional states—their "arms" reach out to the environment, and their identity depends, as Ruth Millikan has stressed, on their evolutionary history.[10]

2. Putnam on naturalism and realism

Hilary Putnam strong refusal of the reductionist "mind-brain type identity view" and his attempts at formulating an adequate non-reductive functionalist view of the mind show that since the early stages of his philosophical career he had been looking for a satisfying non-reductive form of realistic naturalism.[11] What he aimed at was a view that, while deeply

[8] H. Putnam, "From Quantum Mechanics to Ethics and Back Again", in *Philosophy in an Age of Science*, cit., pp. 29–30.

[9] See H. Putnam's *Representation and Reality,* Cambridge (MA), MIT Press, 1988, and Id., *Self-Portrait,* in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind,* Blackwell, Oxford 1993, pp. 507– 513.

[10] H. Putnam, "From Quantum Mechanics to Ethics and Back Again", in *Philosophy in Age of Science*, cit., pp. 51–71; quotation p. 51

[11] See H. Putnam, *Naturalism, Realism, and Normativity*, cit.; see also M. De Caro, "Putnam's Philosophy and Metaphilosophy", introduction to that volume, pp. 1–18. Forms of non-realistic naturalism have also been developed (think of Bas van Fraassen's or John Dupre's views, which deny that we can know that the atoms and the other unobservable entities exist).

respectful of the results obtained by the natural sciences (by refusing all supernaturalist interferences), did not assume that, at least in principle, such sciences could explain everything that could be explained.

Putnam's first accomplished attempt to formulate a satisfying non-reductionist form of naturalism was his "internal realism," which took form in 1976.[12] The core of that conception was an epistemic view of truth (truth in idealized epistemic conditions) inspired by C.S. Peirce and Michael Dummett. This view identified truth with justification in idealized epistemic conditions, which Putnam interpreted in a soft way:

> If I say "There is a chair in my study", an ideal epistemic situation would be to be in my study with the lights on or with daylight streaming through the window, with nothing wrong with my eye-sight, with an unfocused mind, without having taken drugs or been subjected to hypnosis, and so forth, and to look and see if there is a chair there.[13]

The main reason for which Putnam developed that view was his refusal of "metaphysical realism" (a view that had attracted him in the previous couple of decades), according to which reality can be completely described in exactly one way and that way precisely and ultimately fixes ontology. It should be noted, however, that even during his internal realism period Putnam never entirely abandoned scientific realism. First of all, he always thought that the theoretical terms of our best scientific practice do refer to real entities, even if such entities are in principle unobservable (i.e., electrons and black holes), and this means that he always rejected all forms of scientific antirealism, such as instrumentalism, conventionalism, operationalism, and relativism. Second, when he moved away from his pre-1976 physicalistically-oriented realism toward internal realism, Putnam was motivated by his desire to shape a satisfying philosophical realism—that is, a realism able to accept simultaneously (1) the approximate and revisable correctness of the scientific worldview and (2) the approximate and revisable correctness of the ordinary worldview, which Putnam then believed (and never stopped believing) was threatened by the reductionist conceptions of reality. In this light, at least since 1976, Putnam rejected all positions that are unable (or worse, do not even try) to do justice, at the same time, to science and to the ordinary view of the world. Arguably, Putnam's painstaking and uninterrupted efforts to shape a version of naturalistic realism able to acknowledge the partial and revisable verisimilitude of both the ordinary and the scientific images of the world is one of his most relevant bequests to the next generations of philosophers.

Then, starting in 1990, Putnam abandoned internal realism; and that happened for two main reasons. First, he realized that the epistemic conception of truth was deeply inadequate. A convincing supporting example

[12] The best presentation of internal realism is in H. Putnam, *Reason, Truth, and History*, Cambridge University Press, Cambridge 1981.

[13] H. Putnam, *Realism with a Human Face*, Harvard University Press, Cambridge (MA) 1990, p. vii.

of why truth is not epistemically constrained is a conjecture such as "There is no life outside the earth" – which may well be true but, if it is so, it would be unverifiable even in ideal epistemic conditions. In abandoning the epistemic view of truth, however, Putnam realized that he did not need to go antirealist in order to refuse the dogmatic view that he had called "metaphysical realism". His new aim was, in fact, to develop "a modest non-metaphysical realism squarely in touch with the results of science".[14]

The second reason that convinced Putnam to abandon internal realism was the so-called "no-miracles argument", which he had developed in 1973, but whose relevance regarding the issue of realism he fully appreciated much later. This argument is based on the idea that the only way of accounting for the great explanatory and predictive success of the best theories of modern science is to acknowledge that these theories are true (or approximately true) in regard to the natural world and that they refer to real entities, even when those entities are unobservable. From the point of view of antirealism, on the contrary, the fact that science works so well in offering comprehensive explanations and extremely precise predictions of observable phenomena is an inexplicable mystery, if not a sheer miracle.[15] Consequently, according to Putnam, we should consider our best scientific theories as approximately true and the entities such theories refer to as real. Unsurprisingly, antirealists have tried to attack the miracle argument in various ways, but in my view, Putnam and others have responded to those arguments in satisfying ways.[16]

Therefore, according to Putnam – apart from his internal realist period – scientific theories can be true (or approximately true) even in case we cannot ever verify them. However, even if he was a scientific realist, Putnam refused the strict naturalist view – and endorsed the liberal naturalist one – for two main reasons. First, because of the phenomenon he called "conceptual relativity," which means that some theories can be *cognitively equivalent*, even if prima facie they appear to be incompatible. (This phenomenon could less equivocally be called "descriptive equivalence," since the original term may suggest a connection with relativism and antirealism that is entirely inappropriate.) As Putnam convincingly argued, in some scientific fields such as mathematical physics, this phenomenon is ubiquitous:

> To take an example from a paper with the title "Bosonization as Duality" that appeared in Nuclear Physics B some years ago, there are quantum mechanical schemes some of whose representations depict the particles in a system as bosons while others depict them

[14] H. Putnam, *Ethics without Ontology*, Harvard University Press, Cambridge (MA) 2004 p. 286, n.1.

[15] During his internal realism years Putnam defended both the no-miracles argument and the epistemic view of truth: he solved this obvious tension by abandoning the latter view in 1990 (analogously, as we have seen, between 1960 and the mid-1980's he tried to reconcile computational functionalism with semantic externalism, until he realized that they were incompatible).

[16] H. Putnam, "Corresponding to Reality", cit.

> as fermions. As their use of the term "representations" indicates, real live physicists—not philosophers with any particular philosophical axe to grind—do not regard this as a case of ignorance. In their view, the "bosons" and "fermions" are simple artifacts of the representation used. But the system is mind-independently real, for all that, and each of its states is a mind-independently real condition, that can be represented in each of these different ways. And that is exactly the conclusion I advocate...[These] descriptions are both answerable to the very same aspect of reality...they are "equivalent descriptions".[17]

The second reason for which Putnam refused strict naturalism is more interesting for our purposes. This is the fact that, in his view, the ontology of the world cannot be limited to the entities and properties described by physics:

> I do indeed deny that the world can be completely described in the language game of theoretical physics; not because there are regions in which physics is false, but because, to use Aristotelian language, the world has many levels of form, and there is no realistic possibility of reducing them all to the level of fundamental physics.[18]

One of Putnam's favorite examples in this sense was that, depending on what our interests are, we can correctly and usefully describe a chair in the alternative languages of carpentry, furniture, design, geometry, or etiquette. Each of those descriptions is useful in its specific way, without being reducible to any of the others. Moreover, there is no fundamental and unifying theory of what being a chair is, so to speak. And this is true of a vast amount of entities (possibly of all of them, with the exception of the entities of microphysics), since they can all be described in different ways; and this is not just because of conceptual relativity, but also because things have different properties that belong to different ontological regions. A poem is real, for example, but certainly its properties cannot be accounted for by any natural science; or moral (or immoral) actions exist, but not in the ontological region of physics – even if, for Putnam, being a poem or being a moral action are properties that globally supervene on physical properties.

According to Putnam, the old ontological project of providing a unified inventory of the universe, which would supposedly encompass the referents of all possible objective statements – a project of which contemporary metaphysical realism is a very clear expression – has made us wandering in Cloud Cuckoo Land for too long.[19] And this means, for Putnam, that Ontology with a capital "o" is a dead project. However, another form of ontology (one with a lowercase initial) is still possible, i.e., the search for the entities our best theories and practices commit us to. The latter project, however, cannot be carried out if one is driven by the ideological bias that there has to be one, and only one, true

[17] *Ibid.*, p. 64.

[18] *Ibid.*, p. 65.

[19] H. Putnam, *Ethics without Ontology*, cit.

theory of the world; nor can it be carried out without noticing that there are different mutually irreducible ontological levels. And it is a pragmatic question which level is relevant to a particular discursive practice. In this light, Putnam's liberal naturalism incorporated *causal pluralism* (in his view causation and explanation are inextricably interconnected notions)[20] and *the refusal of the fact-value dichotomy* (since, in his view, values and normativity are ubiquitous).[21]

Putnam's view of the mind changed in parallel with his attempts at defining a satisfying form of non-reductive naturalism. In this sense, Putnam remained a functionalist in a broad sense since he continued to understand the mind in terms of its "functions", both internal and external, although not now characterized in computational terms. His later conception of the mind as a "structured system of object-involving abilities",[22] however, built upon his long-standing commitment to semantic externalism by taking seriously that there is no interface between the mind and world in perception or conception:

> The identification of naturalism with such "reduction programs" as the program of reducing the intentional to the non-intentional, or dispensing with intentional and normative notions entirely is a mistake, and I have been explaining how that mistake led me, at one time, to abandon the very realist intuitions with which I started. Some naturalists are reductionists, to be sure, and reduction programs *have* sometimes succeeded, but counting oneself as a naturalist does not require one to subscribe to reduction programs that are, as far as we can now tell, utterly unrealistic. The liberalized functionalism I advocate is an antireductionist but naturalist successor to the original, reductionist, functionalist program. For a liberalized functionalist, there is no difficulty in conceiving of ourselves as organisms whose functions are, as Dewey might have put it, "transactional", that is environment-involving, from the start.[23]

According to the later Putnam, then, a feasible liberal functionalism about the mind has to be conceived in the context of a general liberal naturalism that reconciles what science tells us about the world with the irreducibility of the intentionality of the mental. And I think that this idea is one of the most important ones that this great philosopher left us with.[24]

[20] H. Putnam, *The Threefold Cord. Mind, Body, and World*, Columbia University Press, New York 1999, pp. 137–150.

[21] H. Putnam, *The Collapse of the Fact/Value Dichotomy and Other Essays*, Harvard University Press, Cambridge (MA) 2002, and Id., "The Fact/Value Dichotomy and Its Critics", in *Philosophy in an Age of Science*, cit., pp. 283–298.

[22] H. Putnam, "Replies", *The Philosophy of Hilary Putnam*, monographic issue of *Philosophical Topics*, 2, 1 1992, p. 256.

[23] Putnam, "Corresponding with Reality", cit., pp. 82–83.

[24] I am greatly indebted to Hilary Putnam for the innumerable conversations I had with him in regard to the issues discussed in this article.

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

1

FILOZOFSKI godišnjak = Belgrade philosophical annual / editor Slobodan Perović. – God. 1, br. 1 (1988)– . – Belgrade : Institute of Philosophy, Faculty of Philosophy, 1988- (Belgrade : Službeni glasnik). – 24 cm

Godišnje. – Glavni stvarni naslov od br. 28 (2015) Belgrade philosophical annual. – Tekst na engl. jeziku.

ISSN 0353-3891 = Filozofski godišnjak

COBISS.SR-ID 15073792