A short history of theories of intuitive theories

Johannes B. Mahr (Harvard) & Gergely Csibra (CEU)

To appear in J. Gervain, G. Csibra, K. Kovács (Eds.), ***A Life in Cognition: Studies in Cognitive Science in honor of Csaba Pléh****. Springer.*

Abstract

Intuitive theories are sets of integrated concepts and causal laws that people adopt to comprehend, explain, and predict certain phenomena they encounter in the world. These theories are 'intuitive' because they are thought to drive our intuitions about how the physical and biological world, the mental life of people, and the society we live in work, without meeting the standards of explicit scientific theorizing. The proposal that people adopt such theories has been around at least since the 1970s. However, how psychologists think about intuitive theories has been changing since they have been first proposed. In this chapter, we provide a short overview of the approaches to the function of intuitive theories and belief-forming systems more generally. While early characterization of intuitive theories emphasized their epistemic function, later attempts took an evolutionary view, claiming that they serve adaptive functions that are not always aligned with the goal of accurately tracking environmental states. A recent twist in this story is the proposal that shared intuitive theories may also serve social functions by providing a 'theoretical common ground' on which people interpret unobservable entities, such as memories, character traits, entitlements, and obligations. Such shared theories might be essential for social coordination via communication.

One of the fundamental questions of cognitive science is how human cognition manages to produce adaptive behavior and accurate models of its environment. A popular answer has been that one way human cognition solves this challenge is by building 'theories' about the causal relationships obtaining in the world. Such 'intuitive theories' are suggested to be the main drivers of the kinds of beliefs we form; they are a central part of the human mind's 'belief-forming systems'. What is the function of such belief-forming systems? Do we form beliefs to always accurately model the world, or are there cases in which our mind is set up in a way so as to adaptively misrepresent our environment? In this chapter, we will explore the history of the debate about this question through the lens of the literature on intuitive theories. After providing a short introduction to the concept of 'intuitive theories', we will go through what we perceive to be the main stages in the debate about the function of such theories.

This debate has mainly centered on the question to what extent our belief-forming systems have been set up to be 'truth-approximating'. While in the early days of cognitive science intuitive theories (and belief-forming systems more generally) were thought to be relentlessly truth-tracking, later approaches – inspired by advances in evolutionary psychology – pointed out that considerations other than truth might also have impacted the evolved design of these systems. On the one hand, considerations about cognitive economy and learning under uncertainty have been shown to have the potential to bias belief formation. On the other hand, strategic social considerations in communication might make it adaptive to form wrong beliefs under certain conditions. After providing a short review over the different perspectives on the relation between intuitive theories and truth, we will contribute to this debate by proposing that intuitive theories might serve coordinative social functions that are different from both strategic social and other individualistic considerations. To illustrate this idea, we will discuss examples of how intuitive theories might serve coordinative functions from the domains of episodic memory and intuitive personality psychology.

1. Intuitive theories

How does a helicopter fly? Even though you might think to have a rough idea of how this works, your everyday understanding of helicopters is, most likely, quite impoverished (Keil, 2003). The fact that our everyday understanding of the world largely does not do justice to its deep complexity has arguably given rise to collective enterprises and

institutions of knowledge generation in various human cultures, codified, for example, in scientific methods.

The complexity and difficulty of explaining how even simple processes in the real-world function stands in stark contrast to our ability to competently navigate the world on a day to day basis. For example, people can plan parabolic trajectories of objects they throw or get out of the way of approaching objects without being able to explain the complex physical and geometric relationships underlying these phenomena. What explains such striking competence 'without understanding'? And how do humans acquire this competence in development?

One answer to these questions that has been particularly popular in the cognitive sciences relies on the idea that humans are 'intuitive scientists' who employ 'intuitive theories' for dealing with phenomena in different domains, such as physics, biology, psychology, sociology, or even economics (Karmiloff-Smith & Inhelder, 1975; Wellman & Gelman, 1992; Gopnik & Wellman, 1994). Humans might be able to navigate the world so competently while lacking 'proper' scientific understanding of the underlying processes because their cognitive system builds theories that are similar in many respects to scientific theories. Although these intuitive theories are much shallower than scientific ones (Keil, 2003, 2012), they nonetheless allow for prediction and explanation of, as well as intervention on, the world (Gerstenberg & Tenenbaum, 2017).

What are intuitive theories? Like scientific theories, intuitive theories are integrated systems of causal laws and concepts applying to phenomena in a certain domain. Intuitive theories postulate 'hidden', unobservable entities (e.g., 'forces', 'germs', 'beliefs', 'traits', 'kinds' etc.) that causally act in specific ways to produce observable phenomena in the world. Like scientific theories, intuitive theories are thought to make predictions about the world that can then be used to revise, support, or extend the theory. Moreover, intuitive theories do not only interpret and react to evidence, they also determine what phenomena are relevant, that is, what counts as evidence in the first place. Unlike scientific theories, however, the assumptions and the concepts of intuitive theories are not always explicitly reportable, and the inferences we make on their basis are not always conscious (Uleman, Adil Saribay, & Gonzalez, 2008). Instead, the contents of our intuitive theories commonly express themselves in what we take to be 'obvious', 'intuitive', or 'not in need of explanation'.

Note that one should draw a distinction between an intuitive theory and the beliefs that result from adopting such a theory. In a given context, an intuitive theory might make a certain belief more likely (or 'intuitive') than another. Nonetheless, the beliefs one forms on the basis of a theory are not identical to the theory itself, nor are such theories the only sources of belief formation. Given that intuitive theories are part of human 'belief-forming systems', one way to look at the function of these theories is through the lens of the beliefs they produce.

The notion of an intuitive theory in itself can be construed to be compatible with both nativism and empiricism. From a nativist perspective, one can point out that human beings (and other animals) likely come into the world equipped with some conceptual primitives and the necessary equipment to learn from evidence on their basis (Boyer, 2000; Spelke & Kinzler, 2007). From an empiricist point of view, a specific capacity for theory building might allow us to posit new conceptual entities (Gopnik, 2003), and these capacities may be combined to provide an explanation of genuine conceptual development (Carey, 2009).

The seminal study of Heider & Simmel (1944) illustrates the power of intuitive theories particularly well. In their experiments, participants were presented with a short video clip of interacting geometric shapes. When asked to describe the video clip, most participants interpreted the interaction of the shapes in 'mentalistic' terms, attributing mental states such as beliefs, desires, memories, and plans to the shapes. If participants had simply described what they were seeing, they would have merely referred to the movement patterns of each shape. However, the participants interpreted the shapes as representing agents, and their intuitive psychology recruited concepts like 'belief' and 'desire' to account for the behavior of these agents. These mental state concepts allowed the participants to explain the behavior of the shapes in a way that went decidedly beyond the perceptual evidence. This shows not only that participants had some intuitive theory of how agent-like behavior could be explained in terms of abstract mental state concepts. It also shows that the theoretical entities that make up intuitive theories have to be abstract enough to allow one to 'go beyond' the evidence.

The Heider & Simmel example comes from the domain of intuitive psychology (Premack & Woodruff, 1978; Wellman, 1990; Gopnik & Wellman, 1992). However, the notion of intuitive theories has been extraordinarily productive in explaining people's competencies in a wide range of other domains as well. People are able to think intuitively about the physical world; they are 'intuitive physicists'. Intuitive physics concerns intuitions about

questions such as what kind of arrangements of physical objects are stable or unstable, what the future trajectories of moving objects will be, or about the of relative mass of objects based on their behavior in a collision event (McCloskey, 1983; Kubricht, Holyoak, & Lu, 2017).

Similarly, humans also seem to be in possession of an 'intuitive biology', that is, of a theory of living things and biological processes (Carey, 1985). Across cultures, people seem to treat plants and animals as special kinds of objects different from other parts of the physical world. Just as wide spread seems to be the intuition that living kinds can be arranged according to hierarchical taxonomies and that members of the same living kind share internal, essential features that go beyond their perceivable attributes (Atran, 1998; Medin & Atran, 2004). These assumptions of 'intuitive biology' emerge early in development (Carey, 1985; Keil, 1986; Gelman & Markman, 1986).

More recently, research on how people conceptualize the structure of the social world has been interpreted to suggest that they have an 'intuitive sociology' (Hirschfeld, 1996; Rhodes, 2013). People readily categorize others as belonging to different social groups and assign traits to, as well as modify their behavior towards them based on such perceived group membership (e.g. Taifel, 1978). In fact, one of the cognitive bases for racism seems to be people's tendency to monitor their social environment for signs of group membership and coalitional affiliations (Kurzban, Tooby, & Cosmides, 2001). Even children expect the social world to be composed of 'social kinds', each of which has essential traits that define them (similar to biological kinds) beyond the observable features of their individual members. Children further expect group membership to come with obligations of their members towards other group members. Infants already look for and are sensitive to signals of group membership (such as language) in their social surroundings and form expectations about social affiliation based on such signals (Liberman, Woodward, & Kinzler, 2017). Once group membership has been identified, preschoolers have been shown to use this information productively in inference, for example, to produce causal explanations for the behavior of group members (Rhodes, 2014).

Finally, another domain in which the concept of an intuitive theory has gained some traction recently is what might be called 'folk economics': the intuitions that people bring to questions regarding the exchange of goods and the nature of value. Boyer & Petersen (2018) have proposed that there are cross-culturally stable intuitions about topics such as the benefits of international trade, the effect of immigration on labor markets, and the viability

of social welfare programs that might be the outcome of automatic inference systems evolved as adaptations to life in small-scale societies. In particular, it seems that people have strong intuitions that trade must be a zero-sum game in which one party wins while the other loses (Johnson, Zhang, & Keil, 2020).

This short summary illustrates that the notion of intuitive theories can make intelligible how knowledge and reasoning is organized according to different domains. However, the target phenomena of this article are not intuitive theories themselves but the scientific theories that postulate the existence of such intuitive theories. In spite of the extraordinary productivity of this notion in cognitive science, ideas about why people should be equipped with such theories (and, correspondingly, their relation to 'truth') has gone through considerable change since the concept has first been proposed. While initially the analogy between intuitive and scientific theories was taken to imply that intuitive theories should be 'truth-tracking' (i.e., geared towards optimizing accuracy), later accounts emphasized their adaptive value allowing for 'useful fictions'. In this chapter, we review the different ways the function of intuitive theories has been conceived, and shortly introduce, what we perceive to be, a new take on this topic.

2. Intuitive theories might function to maximize the accuracy of our beliefs

Why do we have intuitive theories? Above, we mentioned a typical answer to this question: intuitive theories allow us to generate explanations and predictions of, and interventions on, the world. In this vein of thinking, many (e.g. Quine, 1977; Millikan, 1984; Fodor, 1983; 1985; 2001; Dennett, 1987) have argued that belief-forming systems must be strongly constrained by truth. After all, the argument goes, "[c]reatures inveterately wrong in their inductions have an unlikely but praiseworthy tendency to die" (Quine, 1977, p. 13). If our theories or beliefs are to support action, prediction, and explanation, they can do so only to the extent that they are reliably accurate.

Such epistemic optimism has been in large part inspired by a combination of two ideas. On the one hand, defenders of the 'epistemic functions' view commonly hold that *successful action* requires *accurate representation*. Jerry Fodor (2008, p. 12) in particular has notoriously argued that "thought is prior to action (because acting requires planning and planning is a species of reasoning)." Trying to catch a ball with an intuitive notion of physics that doesn't reliably produce an accurate prediction of its trajectory will only lead to reliably being hit in the face. Therefore, even though our intuitive theories might sometimes

malfunction to get things wrong, their fundamental function is, on this view, to get things right about the world.

On the other hand, some authors defending the 'epistemic functions view' have drawn optimism from evolutionary considerations. Even though, as mentioned above, the notion of an intuitive theory is compatible with both nativism and empiricism, any empiricist wishing to adopt it will arguably have to concede that at least the theory building equipment will not itself be learned but rather an outcome of evolutionary selection. As such, intuitive theories should be truth-tracking because truth is the currency of survival, and, therefore, natural selection must have to some extent optimized minds to track their environment accurately. Again, this is not to say that our theories aren't sometimes wrong under some descriptions. For example, we seem to have an intuitive notion of physics akin to medieval impetus theory which is surely wrong according to the standards of modern scientific physics (McCloskey, 1983). However, owing to natural selection, our theories will nonetheless be 'designed' to be as accurate as possible given the evidence.

This last point requires unpacking because anyone who believes that the human mind is fundamentally directed towards truth will have to explain how it is that we so often get things so fundamentally wrong. There are, after all, people who believe that the Earth is flat, that Hilary Clinton is running a child prostitution ring out of the basement of a pizzeria, and that vaccines cause autism. In order to accommodate the many circumstances in which people get things wrong, a defender of the 'epistemic functions' view will have to point to *malfunction*. On this view, a misbelief will be the product of the failure of the 'normal workings' of our cognitive apparatus. This is not the place to go into what the best way to think of such 'normal workings' should be (but see e.g. Millikan, 1993, on this point). Suffice it to say that if intuitive theories are thought to primarily serve epistemic functions, they should produce overall accurate beliefs in their normal operation.

3. Intuitive theories might function to solve the trade-off between accuracy and computational efficiency

To many, the idea that the only way the human cognitive system could produce inaccurate beliefs is due to malfunction seems unsatisfactory. There are, after all, many evolutionary considerations why a system that is under selection pressure for producing accurate representations could nonetheless develop to produce 'wrong' beliefs — while still being adaptive. Intuitive theories, to the extent that they are a product of evolution, might

well produce wrong beliefs in their *normal operation*. In the words of Paul Bloom (2004; p. 222-223): "All other things being equal, it is better for an animal to believe true things than false things; accurate perception is better than hallucination. But sometimes all other things are not equal." On this view, intuitive theories don't have to be accurate to be useful or adaptive, they only have to be accurate *enough*. Note that this perspective is not necessarily in disagreement with the proposal that belief-forming systems have primarily epistemic functions. The disagreement is rather about some other considerations that may take precedent over accuracy in how our cognitive system is shaped by natural selection.

First among the ways in which "all other things are sometimes not equal" causing a system to produce wrong beliefs is arguably cognitive economy: accurate beliefs might often be computationally too expensive to be worth the effort. Moreover, selection might not 'optimize' systems to produce true beliefs under all circumstances, but the accuracy of our judgments might instead be highly 'bounded' to specific contexts (Simon, 1956). Research in the vein of 'bounded rationality' has thus pointed out that biased judgment and false beliefs can be the outcome of the entirely normal workings of a cognitive system. In particular, the 'heuristics and biases' (Tversky& Kahneman, 1974; Kahneman, Slovic, & Tversky, 1982 and 'fast and frugal heuristics' (Gigerenzer & Selten, 2002; Gigerenzer & Todd, 1999) research programs have shown that the human mind can produce biased judgments in response to a trade-off between strict accuracy, computational efficiency, and environmental specificity. Heuristics, i.e., simple decision rules for complex problems, are thought to solve such trade-offs by generating correct intuitions in specific circumstances without going through computationally expensive algorithms that would be required for inevitably arriving at the correct answer under all circumstances.

While heuristics are commonly not presented in terms of theories, theories, too, can have heuristic qualities. For example, one way to make sense of the characteristic mistakes people make in physical judgments, is to attribute to them a 'heuristic model' of physical quantities and relationships (Kubricht, Holyoak, & Lu, 2017). A heuristic way to compute the relative mass of two colliding objects, for example, would be to compare whether the post-collision velocity of object A is greater than that of object B. If yes, object A will be taken to be lighter, if no, object B will be taken to be lighter (Gilden & Proffitt, 1994).

In other domains, too, it has been proposed that ontogenetic and phylogenetic development might produce intuitive theories that can produce incorrect beliefs in their normal operation. Most radically, Churchland (1981) has defended an entirely eliminitavist

position with regard to intuitive psychology: mental state reasoning might be an evolved solution to explain, predict, and interact with others without describing anything 'real' about their minds. Dennett (1987) can also be read as arguing that intuitive psychology (or as he calls it 'the intentional stance') amounts to nothing more than a heuristic calculus for predicting behavior without strictly picking out 'real' entities beyond the patterns of behavior that license the predictive inferences we draw from them: "Intentional systems [do not] really have beliefs and desires, but […] one can explain and predict their behavior by ascribing beliefs and desires to them" (Dennett, 1987; p. 7).

4. Intuitive theories might function to maximize the expected fitness value of our beliefs

Even fictionalism about mental states takes the function of intuitive psychology to be epistemic in principle. Views about the fictional character of intuitive psychology do not claim that intuitive psychology is useful *because* it is wrong but rather in spite of it. If there was an equally efficient but epistemically more accurate way to achieve the predictive and explanatory power that intuitive psychology conveys, natural selection would have favored it.

Another line of reasoning, however, argues that inaccurate intuitions and beliefs might be adaptive not in spite of being wrong but *because of it* (McKay & Dennett, 2009). On this view, "natural selection does not care about truth; it cares only about reproductive success" (Stich, 1990; p. 62). Some parts of the human mind might be built in a way so as to produce inaccurate judgments because having an inaccurate judgment in a certain domain might be more adaptive than an accurate one. Natural selection does not shape intuitive judgments to maximize accuracy but rather *expected fitness value* (Cosmides & Tooby, 1987). Expected fitness value, however, can sometimes be at odds with accuracy and, therefore, our judgments can be adaptive on average, *because* they are wrong. In other words, if there is a choice between maximizing accuracy and maximizing expected value, selection will tend to favor the latter.

One source for such arguments has been error-management theory (Haselton, 2007; Haselton & Buss, 2000; Haselton & Nettle, 2006). Whenever in a given domain one kind of error (false negatives, say) is more costly in terms of fitness than another (false positives, say), error-management theory predicts that a system biased towards the less costly form of error should develop. It is, for example, more costly to underestimate the danger posed by snakes than to overestimate it. While the former can lead to death or severe injury, the

latter will rarely if ever have such harsh consequences. Because our judgments about the danger of snakes commonly happen under uncertainty, from the perspective of error-management theory, selection should thus have favored a system that errs on the side of caution and overestimates the danger of snakes. This line of reasoning has been applied in the explanation of a variety of phenomena such as the perception of approaching sounds, the perception of dangerous animals and people, the perception of sexual interest, agency detection, and many others (Haselton & Nettle, 2006, Haselton, 2003, Barrett, 2000).

5. Intuitive theories might serve strategic social functions

As long as individual decision-making under uncertainty is concerned, selection will tend to favor the maximization of expected value even at the expense of accuracy. There is however another source of adaptive inaccuracy that might have an impact on the design of our belief-forming systems. As soon as one's own pay-off somehow depends on what other agents believe, and manipulation of these agents becomes possible, there might be strategic social benefits of holding wrong beliefs that might offset the costs of inaccuracy.

On the one hand, holding a belief might make it more likely that others will adopt this belief as well. Thus, if one would benefit from others holding a given wrong belief, it might be beneficial to adopt that belief so as to increase the likelihood of it being adopted by others. In other words, deceiving oneself might have benefits through having an effect on the behavior of others (Trivers, 2000; von Hippel & Trivers, 2011), and this by itself might offset the costs of holding a wrong belief. One example that is often cited in this context is that of 'positive illusions' about one's own abilities and traits (e.g. Heck, Simons, & Chabris, 2018; see also Dunning, 2011). If, by adopting positively biased beliefs about oneself, one makes it more likely that others will also adopt those beliefs, this might offset the costs of being biased.

People should therefore be disposed to adopt certain false beliefs to the extent that the strategic social benefits outweigh their costs. In a similar vein, Mercier & Sperber (2011; 2017) have argued that human reasoning has been shaped to show a 'my-side' or 'confirmation' bias in order to make it more likely to produce reasons supporting one's own beliefs rather than to contradict these beliefs. Such a bias would be adaptive in the context of argumentation where skeptical interlocutors are not convinced based on trust alone but require justification.

On the other hand, false beliefs can have social benefits not in spite of the fact that they are otherwise costly but because of it. Holding a false belief (and producing behaviors that such beliefs trigger) might be a costly signal communicating to others that one is one is ready to bear the cost of such belief in order to belong to a particular group. Particularly, if a given belief precludes group membership in other groups, it might be effective in signaling that one has "burned one's bridges" and therefore is truly committed to one's in-group. For example, religious beliefs and ritual behaviors (Irons, 2001; Sosis & Alcorta, 2003), and the absurd beliefs mentioned above ("Flat Earthers", "Pizza Gaters", etc., Mercier, 2020) have been analyzed in these terms. It is worth noting at this point, however, that beliefs serving as costly signals are commonly held not intuitively but rather reflectively (Sperber, 1997): they are held on the basis of abstract reasons and only influence action to the extent required by their function as commitment devices. The extent to which *intuitive* theories can therefore be analyzed as costly signals or commitment devices remains unclear.

More generally, the ideas discussed in this section might merely apply to the strategic social functions of false *beliefs* and not entire *theories*: whenever a given false belief has strategic social benefits, these benefits will often not extend to an underlying layer of concepts and integrated causal laws of an entire domain. As such, if people adopt false beliefs for strategic social benefits, they will most likely not do so based on intuitions generated through their intuitive theories.

6. Intuitive theories might serve coordinative social functions

It is possible that intuitive theories and the beliefs they produce can have strategic communicative consequences either by supporting the manipulation of others' beliefs or by acting as costly signals. These are, however, not the only or even primary social functions intuitive theories might serve. After all, the intuitions that someone brings to a social interaction determine how she coordinates with others. If two people act on the basis of entirely different models of a certain domain, coordinated behavior (of which communication is one instance) will be difficult. For example, the idea that our minds are designed to produce reasons biased in favor of what we already believe (Mercier & Sperber, 2011; 2017) makes sense only to the extent that others share intuitions about what are good and bad reasons and about the norm that assertions need justification in the first place.

Thus, the way individuals model the social world has consequences on what claims they will accept and how they will behave towards each other. As such, some intuitive

theories have important repercussions for the organization of social life, and these effects might drive the development of intuitive theories certainly in ontogenetic, but also potentially in phylogenetic, development. If one developed intuitions that would be fundamentally at odds with how one's social environment took the world to be, one would lose out on a large range of social opportunities. How intuitive theories become coordinated in a society is not a simple question and answering it will certainly require taking into account both historical factors (i.e., cultural evolution), and evolutionary and ecological considerations. Therefore, we will not try to answer this question here. Instead, we will briefly explore two domains of shared intuitions without which the social coordination of obligations, entitlements, responsibilities, and accountabilities would be impossible to regulate in a society. In both cases, it is clear that the underlying theories could produce epistemically invalid beliefs; yet such beliefs may contribute to the coordination and stabilization of shared social facts.

Our first example concerns people's intuitive notion of 'remembering'. It seems that people have an intuitive theory about what constitutes 'real' remembering: the agent in question has personally experienced the remembered event which resulted in a kind of stored 'memory object' (or 'trace'), the retrieval of which recreates the perceptual details of the original event (e.g., Craver, forthcoming; Mahr, 2019; Martin & Deutscher, 1966; Roediger, 1980).[1] In this sense, memory owes its epistemic reliability to the hyper-veridical status that we attribute to perception. Such an intuitive notion of memory explains why claims made on the basis of first-hand evidence (i.e., testimony) *ceteris paribus* seem to be taken to have higher epistemic reliability than beliefs formed on the basis of other sources (Mahr & Csibra, 2020a). However, in contrast to this intuitive characterization of remembering, research on episodic memory has consistently found that remembering is a constructive, highly inferential and malleable process (Bartlett, 1932; Schacter & Addis, 2007).

How can we make sense of the contradiction between the intuitive epistemic immediacy and empirical constructiveness of remembering? Elsewhere, we have proposed that what makes a testimony appear more reliable than other types of information is not its epistemic validity but the very fact that the speaker is willing to take responsibility for its

[1] Note that, as such, claims to remembering seem to be subjected to a higher degree of epistemic vetting than claims to 'knowledge' more generally. While claims to knowledge only require *some kind* of justification, in the case of remembering, the justification has to come in the form of first-hand experience.

content (Mahr & Csibra, 2018; Mahr & Csibra, 2020a). Coordinating beliefs about what happened in the past plays a particularly central role in human social life: most of our social commitments, entitlements, and obligations are ultimately only justifiable through reference to past events (Mahr & Csibra, 2020b). And since what happened in the past can normally be assessed only via testimony, we need common criteria to judge what counts as reliable evidence of past events. The intuitive norms we apply to remembering, therefore, might be due to the fact that humans have to socially regulate who can be given epistemic authority about the past. After all, with such authority comes the power to arbiter over present social realities (think of where eye-witness testimony becomes most important).

Another domain where functions of social coordination might help explain the structure of our intuitions is in the domain of moral psychology. A number of recent approaches to moral judgments have argued that people act as 'intuitive ethicists' (Landy & Uhlmann, 2018; Uhlmann, Pizarro, & Diermeier, 2015), or more specifically, 'intuitive virtue theorists'. According to this view, people have a theory of 'moral traits', and evaluate others according to these traits over and above the moral permissibility of their actual actions. In fact, striking dissociations between people's moral evaluations of specific acts and the person carrying out that act have been found: people seem to take some actions to offer strong evidence of a certain moral character trait without being otherwise ethically objectionable. For example, in a study by Uhlmann, Zhu, and Diermeier (2014), participants perceived the use of a racial slur as being stronger evidence of poor moral character than physical assault even though they judged physical assault itself to be the more blameworthy action.

While there seems to be good evidence for the fact that people attribute moral character traits, it is less clear to what extent these attributions reflect real, stable, situation- and partner-independent behavioral dispositions. There is some disagreement on the extent to which we should think that people indeed possess moral character traits or 'virtues' and 'vices' (e.g. Fleeson et al., 2014; Alfano, 2013). Regardless of the validity of moral trait attributions, however, such attributions themselves might play a role in regulating others' behavior. If people did not have an intuitive theory to the effect that others are endowed with stable 'good' and 'bad' moral character traits that go beyond the ethical permissibility of their individual actions, there would be no basis on which to choose social partners (Martin & Cushman, 2015). If, however, people choose cooperative partners on the basis of moral character, reputational concerns will force everyone to act in ways to appear virtuous

- whether or not this behavior originates from some underlying traits. In other words, the shared intuitive theory that prompts us to tag people with character traits can serve the function of making us behave more virtuously, thereby promoting social cooperation. And crucially, this intuitive theory works only if it is shared, and is expected to be shared, in the community, i.e., if it establishes a theoretical common ground among potential social partners.

7. Conclusion

The question to what extent belief-forming systems in human beings are 'designed' to produce true beliefs has been at the root of many debates in cognitive science. While initially it was taken for granted that natural selection would ensure that our cognitive system should be optimized to produce accurate representations, this optimism has been questioned from a variety of directions. Considerations of context-specificity, computational frugality, adaptive value, and strategic social manipulation have all been shown to be able to enable the selection of cognitive systems that produce false beliefs in their normal operation. We have proposed that beyond these mechanisms, an additional factor might contribute to the emergence of intuitions that are not primarily constrained by the need for accurately representing the world: social coordination. The need for social coordination might explain human intuitions about what constitutes reliable knowledge about the past and why others should have stable moral character traits.

Acknowledgements

Csaba Pléh was instrumental in nurturing in us the attitude that novel ideas in science should always be considered taking into account their roots and historical contexts. He also helped us a lot in developing our theory of remembering (Mahr & Csibra, 2018) and how it is related to intuitive understanding of memory. Beyond him, we also thank Denis Tatone, Kristóf Kovács, and an anonymous reviewer for their valuable comments on an earlier version of this chapter. This work was partly supported by the Advanced Investigator Grant #742231 (PARTNERS) from the European Research Council (ERC) as well as a Mind, Brain, and Behavior Faculty award from Harvard University.

References

Alfano, M. (2013). *Character as Moral Fiction*. Cambridge University Press.

Atran, S. (1998). Folk biology and the anthropology of science: cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*(4), 547-569.

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences, 4(*1), 29-34.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Bloom, P. (2004). *Descartes' Baby*. Random House.

Boyer, P. (2000). Natural epistemology or evolved metaphysics? Developmental evidence for early-developed, intuitive, category-specific, incomplete, and stubborn metaphysical presumptions. *Philosophical Psychology*, *13*(3), 277-297.

Boyer, P. & Petersen, M. B. (2018). Folk-economic beliefs: An evolutionary cognitive model. *Behavioral and Brain Sciences*, *41,* E158.

Carey, S. (1985). *Conceptual Change in Childhood*. MIT press.

Carey, S. (2009). *The Origin of Concepts.* Oxford University Press.

Churchland, P. M. (1981). Eliminative materialism and propositional attitudes. T*he Journal of Philosophy*, *78*(2), 67-90.

Cosmides, L. & Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (p. 277-306). MIT Press.

Craver, C. (forthcoming). Remembering: Empirical and normative. *Review of Philosophy and Psychology.*

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In: M. Zanna and J. Olson (Eds*.), Advances in Experimental Social Psychology* (Vol. 44, p. 247-296). Academic Press.

Fleeson, W., Furr, R. M., Jayawickreme, E., Meindl, P., & Helzer, E. G. (2014). Character: The prospects for a personality-based perspective on morality. *Social and Personality Psychology Compass*, *8*(4), 178-191.

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

Fodor, J. (1985). Précis to the modularity of mind. *Behavioral and Brain Sciences, 8*, 1-42.

Fodor, J. (2001). *The Mind Doesn't Work That Way: The Scope and Limits of Computational*

Psychology. MIT Press.

Fodor, J. (2008). *LOT2: The Language of Thought Revisited*. Oxford University Press.

Gelman, S. & Markman E. (1986). Categories and induction in young children. *Cognition*, 23, 183-209.

Gerstenberg, T. & Tenenbaum, J. B. (2017). Intuitive theories. In: *The Oxford Handbook of Casual Reasoning* (p. 515-548). Oxford University Press.

Gigerenzer, G. & Todd, P. M. (1999). *Simple Heuristics that Make Us Smart*. Oxford University Press, USA.

Gigerenzer, G. & Selten, R. (Eds.). (2002). *Bounded Rationality: The Adaptive Toolbox*. MIT press.

Gilden, D. L. & Proffitt, D. R. (1994). Heuristic judgment of mass ratio in two-body collisions. *Perception & Psychophysics*, *56*(6), 708-720.

Gopnik, A. (2003). The Theory Theory as an alternative to the innateness hypothesis. In *Chomsky and His Critics* (pp. 238-254). John Wiley & Sons.

Gopnik, A. & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (p. 257-293). Cambridge University Press.

Gopnik, A. & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1-2), 145-171.

Haselton, M. G. & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, *78*(1), 81-91.

Haselton, M. G. & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, *10*(1), 47-66.

Haselton, M. G. (2003) The sexual overperception bias: Evidence of a systematic bias in men from a survey of naturally occurring events. *Journal of Research in Personality, 37*(1), 34-47.

Haselton, M. G. (2007). Error management theory. In R. F. Baumeister and K. D. Vohs, (Eds.), *Encyclopedia of Social Psychology, Volume 1*, (p. 311-312). Sage.

Heck, P. R., Simons, D. J., & Chabris, C. F. (2018). 65% of Americans believe they are above average in intelligence: Results of two nationally representative surveys. *PloS one*, *13*(7), e0200103.

Heider, F. & Simmel, M. (1944). An experimental study of apparent behavior. *The American*

Journal of Psychology, *57*(2), 243-259.

Hirschfeld, L. (1996). *Race in the Making*. MIT Press.

Irons, W. (2001). Religion as a hard-to-fake sign of commitment. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment* (p. 292-309). Russell Sage Foundation.

Johnson, S. G. B., Zhang, J., & Keil, F. (2020, April 30). Win-win denial: The psychological underpinnings of zero-sum thinking. https://doi.org/10.31234/osf.io/efs5y

Kahneman, D., Slovic, S. P., & Tversky, A. (Eds.). (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Karmiloff-Smith, A. & Inhelder, B. (1975). "If you want to get ahead, get a theory". *Cognition*, *3*(3), 195-212.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368-373.

Keil, F. C. (1986). The acquisition of natural kind and artefict terms. In W. Demopoulos (ed.), *Language Learning and Concept Acquisition* (p. 133-153). Ablex.

Keil, F. C. (2012). Running on empty? How folk science gets by with less. *Current Directions in Psychological Science*, *21*(5), 329-334.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749-759.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*(26), 15387-15392.

Landy, J. F. & Uhlmann, E. L. (2018). Morality is personal. In: K. Gray and J. Graham (Eds.), *Atlas of Moral Psychology* (p. 121-132). Guilford Press.

Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). The origins of social categorization. *Trends in Cognitive Sciences*, *21*(7), 556-568.

Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review*, *111*(4), 960-983.

Mahr, J. B. (2019). Why do we think what we think about what memory is? Talk presented at Issues in Philosophy of Memory 2, Grenoble, France.

Mahr, J. B. & Csibra, G. (2020a). The effect of source claims on statement believability and speaker accountability. Poster presented at the Budapest CEU Conference on Cognitive Science, Budapest, Hungary.

Mahr, J. B. & Csibra, G. (2020). Witnessing, remembering, and testifying: Why the past is

special for human beings. *Perspectives on Psychological Science*, *15*(2), 428-443.

Mahr, J. B. & Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behavioral and Brain Sciences, 41,* E1.

Martin, J. W. & Cushman, F. (2015). To punish or to leave: Distinct cognitive processes underlie partner control and partner choice behaviors. *PLoS One*, *10*(4).

Martin, C. B., & Deutscher, M. (1966). Remembering. *The Philosophical Review*, *75*(2), 161-196.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122-131.

Medin, D. L. & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review*, *111*(4), 960-983.

Mercier, H. & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57-74.

Mercier, H. & Sperber, D. (2017). *The Enigma of Reason.* Harvard University Press.

Mercier, H. (2020). *Not Born Yesterday: The Science of Who We Trust and What We Believe.* Princeton University Press.

Millikan, R. G. (1984). Naturalist reflections on knowledge. *Quarterly Journal of Philosophy, 65*(4), 315.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice*. MIT Press.

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515-526.

Quine, W. V. (1969). Natural kinds. In: N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (p. 5-23). Springer.

Roediger, H. L. (1980). Memory metaphors in cognitive psychology. *Memory & Cognition*, *8*(3), 231-246.

Rhodes, M. (2013). How two intuitive theories shape the development of social categorization. *Child Development Perspectives*, *7*(1), 12-16.

Rhodes, M. (2014). Children's explanations as a window into their intuitive theories of the social world. *Cognitive Science*, *38*(8), 1687-1697.

Uleman, J. S., Adil Saribay, S., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*(1), 329-360.

Schacter, D. L. & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 773-786.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129.

Sosis, R. & Alcorta, C. (2003). Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology, 12*, 264-274.

Spelke, E. S. & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89-96.

Sperber, D. (1997). Intuitive and reflective beliefs. *Mind and Language, 12*(1), 67-83.

Stich, S. (1990). *The Fragmentation of Reason*. MIT Press.

Tajfel, H. (1978). *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations.* Academic Press.

Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences*, *907*(1), 114-131.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72-81.

Uhlmann, E. L., Zhu, L. L., & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry. *European Journal of Social Psychology, 44*(1), 23-29.

Von Hippel, W. & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(1), 1-56.

Wellman, H. M. & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337-375.

Wellman, H. M. (1990). *Children's Theory of Mind*. MIT Press.