Behavioral and Brain Sciences

cambridge.org/bbs

Rationalization is rational

Fiery Cushman

Department of Psychology, Harvard University, Cambridge, MA 02138
cushman@fas.harvard.edu https://cushmanlab.fas.harvard.edu

Target Article

Cite this article: Cushman F. (2020) Rationalization is rational. *Behavioral and Brain Sciences* **43**, e28: 1–59. doi:10.1017/S0140525X19001730

Target Article Accepted: 19 May 2019
Target Article Manuscript Online: 28 May 2019
Commentaries Accepted: 27 November 2019

Keywords:
rationalization; cognitive dissonance; self-perception; theory of mind; inverse reinforcement learning; habitization; social learning; reflective equilibrium; useful fiction; representational exchange

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 16) and an Author's Response (p. 44). See bbsonline.org for more information.

© The Author(s), 2019. Published by Cambridge University Press

Abstract

Rationalization occurs when a person has performed an action and then concocts the beliefs and desires that would have made it rational. Then, people often adjust their own beliefs and desires to match the concocted ones. While many studies demonstrate rationalization, and a few theories describe its underlying cognitive mechanisms, we have little understanding of its function. Why is the mind designed to construct post hoc rationalizations of its behavior, and then to adopt them? This may accomplish an important task: transferring information between the different kinds of processes and representations that influence our behavior. Human decision making does not rely on a single process; it is influenced by reason, habit, instinct, norms, and so on. Several of these influences are not organized according to rational choice (i.e., computing and maximizing expected value). Rationalization extracts implicit information – true beliefs and useful desires – from the influence of these non-rational systems on behavior. This is a useful fiction – fiction, because it imputes reason to non-rational psychological processes; useful, because it can improve subsequent reasoning. More generally, rationalization belongs to the broader class of representational exchange mechanisms, which transfer information between many different kinds of psychological representations that guide our behavior. Representational exchange enables us to represent any information in the manner best suited to the particular tasks that require it, balancing accuracy, efficiency, and flexibility in thought. The theory of representational exchange reveals connections between rationalization and theory of mind, inverse reinforcement learning, thought experiments, and reflective equilibrium.

> *This fight is over.*
>
> *THE MAN standing there. In the silence. Two unconscious cops at his feet. Blood on his pants. What just happened? How did he do this? And there's THE GUN in his hand. And God, it just feels so natural – checking it – stripping it down – holding it – aiming it – like this is something he's done a million times before….*
>
> *This is something he definitely knows how to do.*
>
> – *The Bourne Identity* (film)

1. Introduction

Jason Bourne is an extraordinary man – a special project of the Central Intelligence Agency. Like a robot, he has been programmed with a vast store of actions that are potentially useful to a clandestine agent. He is fluent in a dozen languages; his gut tells him who to trust and who to fear; he drives like an Italian cabby; he is handy with a gun.

But Bourne faces an extraordinary problem. He has lost his memory, identity, goals, and plans – in sum, his ability to make sense of the world and his own place in it. This problem and his solution drive the plot of *The Bourne Identity*. He must figure out what to believe and what to value by making sense of his peculiar, programmed abilities. The very actions that Bourne performs mindlessly – checking a gun, stripping it, holding it, aiming it – are the clues he uses to rebuild his mind. Jason Bourne learns what to think by seeing what he does.

And, in this respect, he is perfectly ordinary. Each of us faces the same problem every day, and each of us grasps for the same solution. We are never fully certain of what to believe and what to value. But by observing the actions we are programmed to perform, we can draw useful inferences – educated guesses about how the world is and what to want from it. Like Bourne's, ours is a rational project: to reverse-engineer the design principles of our automatic actions. Mercifully, for us, the stakes are usually lower. Perhaps that is why it's so fun to watch Jason Bourne: His life is ours, just more so.

1.1. Rationalization

Rationalization takes an action that has already been performed and then concocts the beliefs or desires that would have made it rational. It is, therefore, exactly the opposite of rational action[1] (Fig. 1). Rational action begins with beliefs and desires and then deduces the optimal

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

action to perform – the one that maximizes desires, conditioned on beliefs. If you believe that a man threatens your life, if you want to live, and if you think he can only be stopped with a bullet, then it is rational to shoot him. Rationalization turns this process on its head: First, you shoot a man, and from this you conclude that he threatened your life.

Sensibly or not, people rationalize all the time. Among psychologists, it is one of the most exhaustively documented and relentlessly maligned acts in the human repertoire. Classic topics such as cognitive dissonance (Festinger 1962), emotion misattribution (Schachter & Singer 1962), appraisal theory (Arnold 1960; Lazarus 1982), self-perception (Bem 1967; Nisbett & Wilson 1977), and confabulation (Gazzaniga 1967) all have rationalization at their heart. More peripherally, it supports confirmation bias (Nickerson 1998), system justification (Jost & Banaji 1994), motivated reasoning (Kunda 1990), culpable control (Alicke 2000), hindsight bias (Christensen-Szalanski & Willham 1991), immune neglect (Gilbert et al. 1998), and more.

Some cases of rationalization are easy to explain. Perhaps we are merely attempting to explain our own behavior, inferring its obscure causes (Bem 1967). This occurs when, for instance, you notice your furtive glances and flushed cheeks and exclaim to yourself: "I'm falling in love!" (Dutton & Aron 1974; Schachter & Singer 1962). Other times, we are hoping to convincingly recast our behavior in a more favorable light ("I ate that last cookie so that nobody else would feel awkward about it!"; Mercier & Sperber 2011; Tedeschi et al. 1971; von Hippel & Trivers 2011).

But other cases of rationalization are much harder to explain. In these hard cases people don't just tell a story, they actually make themselves believe it (Brehm 1956; Sharot et al. 2010; Vinckier et al. 2019). In one experiment, for instance, participants were tricked into believing they had made a subliminal choice between two vacation destinations, such as Thailand and Greece. People duped into thinking they chose Greece actually began to like it more, while people who thought they had chosen Thailand showed the opposite preference change (Sharot et al. 2010). Similarly, people believe that a lottery ticket is more likely to win as soon as they have bought it (Langer 1975), or that a horse is more likely to win its race as soon as they have bet on it (Knox & Inkster 1968). These cases are hard to see as anything but gross errors. You are supposed to choose Thailand because you preferred it, or a bet on horse because of its odds. How, then, could those choices justify increasing your preference for Thailand or your belief in the horse's odds? These cases seem stubbornly irrational. Why do we drink our own Kool-Aid?

Current theories of rationalization explain how it works, identifying the underlying psychological mechanisms. For instance, the theory of cognitive dissonance posits that we revise our preferences (for Thailand) and beliefs (in a horse's odds) because we are motivated to reduce dissonance between thought and action. But these theories mostly fail to explain why the brain would contain such mechanisms in the first place (but see Mercier & Sperber 2011). In other words, why would natural selection favor a "dissonance reduction motive"? Classic accounts of rationalization are vague – even silent – at this "ultimate" (Tinbergen 1963) or "computational" (Marr 1982) level of analysis.

Fiery Cushman is the John L. Loeb Associate Professor of Social Sciences in the Department of Psychology at Harvard University. He is a recipient of the Distinguished Scientific Award for Early Career Contribution to Psychology from the American Psychological Association, the Stanton Prize from the Society for Philosophy and Psychology, and the Theoretical Innovation Award from the Society for Personality and Social Psychology. His research centers on human decision making and moral judgment.

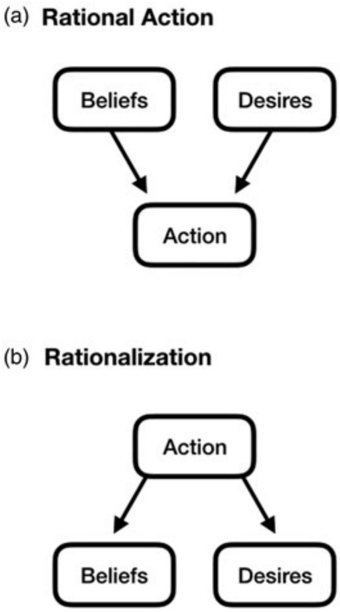

Figure 1. The relationship between (a) rational action and (b) rationalization.

To address this challenge, it helps to return to one of most basic insights of psychological research: Our behavior is influenced by many psychological processes that are (1) unconscious, (2) non-rational, and yet (3) biological adaptive. For instance, our behavior is influenced by instincts, habits, and conformity to social norms (Fig. 2a). Rationalization, then, may be a mechanism for extracting valuable information from these adaptive choices and then allowing it to influence the network of beliefs and desires that support reasoning (Fig. 2b).

According to this view, rationalization is not merely designed to infer the underlying causes of our behavior for the sake of explanation (Bem 1967). It is not, for instance, designed to discover our unconscious reasons: hidden beliefs and desires. Rather, it constructs new beliefs and desires where none had existed, to extract information from the *non-rational* processes that influence our behavior. In other words, just as Jason Bourne has been programmed by the CIA with a host of useful reflexes, we have all been programmed: by natural selection, by habit learning, by social learning, and so forth. Thus, just as Bourne can observe his automatic behaviors and extract useful information, so can we.

A simple example illustrates the basic idea. Suppose that an infant crawls to high point and then pulls back from the edge by instinct (Gibson & Walk 1960). This action does not reflect a belief that heights are dangerous, or the desire to avoid falling; rather, the infant pulls back from the edge by instinct alone (Gendler 2008). But, having performed this action, rationalization seeks to learn from it – first concocting beliefs and desires that could have produced it, and then adopting them. For instance, infants might conclude that heights are dangerous, or adopt the desire to avoid them. Thus infants do not infer the actual beliefs

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

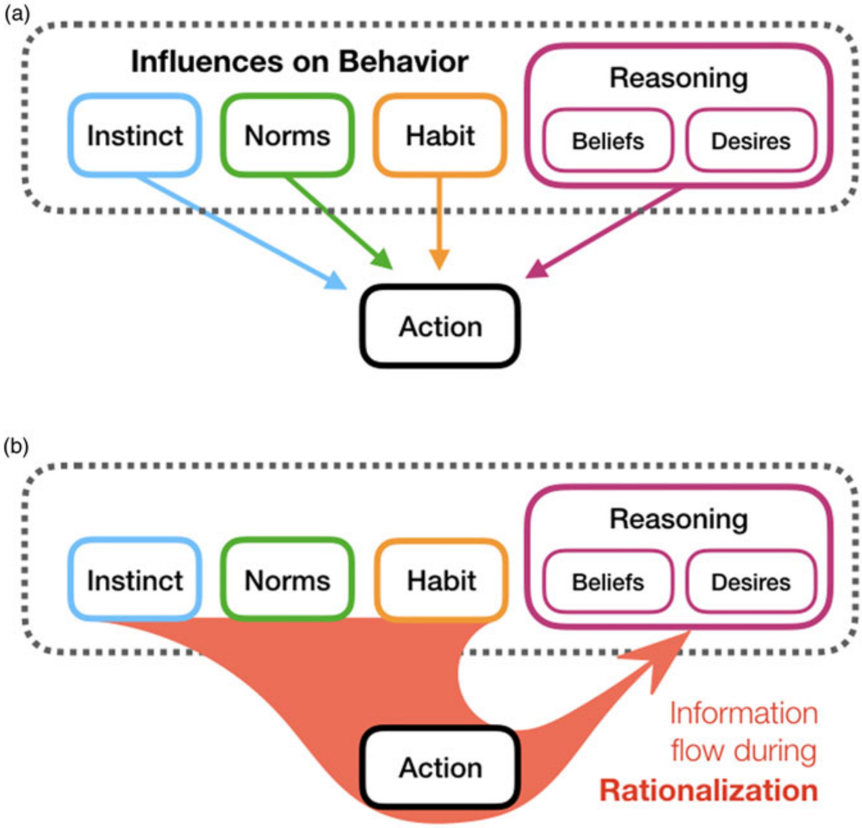

Figure 2. Human behavior is influenced by (a) multiple adaptive processes. (b) Rationalization is a method of extracting information from non-rational processes and making it available to and useful for future reasoning.

or desires that guiding their actions; rather, they construct new ones, imputing false mental states to an instinctual action. Nevertheless, their new beliefs and desires are adaptive – precipices are dangerous, and you should avoid them. This is no accident: Our instincts embody the hard-won lessons of natural selection. They are, therefore, a rich source of information for your rational mind.

As adults, of course, we rarely rationalize in situations as simple as this. For one thing, our behavior is usually the product of multiple influences, not a "pure" effect of reflex, habit, reasoning, and so forth. For another thing, our minds are not blank slates bereft of prior knowledge and preferences. These facts make adult rationalization more complex, but no less adaptive. Insofar as non-rational processes exert *some* influence on our behavior, and insofar as that influence is adaptive, we can extract useful information by adopting the beliefs and desires that would have made our actions rational.

Rationalization, then, is "rational" in two senses. First, it is adaptive. This doesn't guarantee that it always benefits a person in every particular case; to the contrary, many psychologists have made their careers by brilliantly illustrating the ways in which it can fail. Like any process that is generally adaptive, it will be occasionally be maladaptive. But on average, over time, it pays.

Second, at a more specific level, rationalization approximates inverse planning models of mental state inference (Baker et al. 2009; Ng & Russell 2000). Cast as a form of Bayesian inference, inverse planning is sometimes regarded as a rational cognitive process. Of course, the particular mechanisms we use to rationalize our behavior are unlikely to conform to a normative, rational standard (such as Bayesian inference) in every detail. Nevertheless, the basic structure of rationalization can be understood as an approximation of something rational.

Our first two goals are to review existing theories of rationalization and to contrast these with the present account. Finally, this article presents a theory of representational exchange situating rationalization in a broader framework. Representational exchange describes the flow of information between distinct control systems (reasoning, habits, instincts, and norms) to facilitate efficient, adaptive choice. This clarifies the overarching adaptive rationale that unifies rationalization with many other forms of representational exchange and highlights its connections to inverse reinforcement learning, habitization, theory of mind, social learning, thought experiments, and the philosophical pursuit of reflective equilibrium.

2. Existing accounts of rationalization

The psychological literature on rationalization is large and varied. Currently, three basic approaches dominate: (1) cognitive dissonance and consonance, (2) self-perception, and (3) persuasion and impression management. These theories each differ from the current proposal, but do not necessarily compete with it. For one thing, different kinds of rationalization may occur in different contexts. More importantly, existing theories mostly describe the mechanisms of rationalization, whereas the present theory addresses its function. These levels of explanation are often complementary (Marr 1982; Tinbergen 1963). The goal, therefore, is not adjudicate between theories, marshaling data for some and against others. Rather, it is to clearly present each

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

theory, and then to consider their points of convergence and divergence. This discussion focuses squarely on the rationalization of *action*, touching only superficially on other kinds, such as rationalizing one's beliefs or attitudes.

2.1. Dissonance and consonance

Rationalization is usually explained by positing a desire for consonance between thought and action. The simplest model of this kind, balance theory (Heider 1958/2013), posits that people want to achieve harmony among their attitudes. Thus, if you like your spouse and your spouse likes guacamole, you will tend to acquire a taste for guacamole (or, in theory, a distaste for your spouse). Unfortunately, this model's generality is also its Achilles' heel: It is easy to come up with compelling counterexamples and hard for the theory to explain them away. As Festinger (1999) quipped, "I like chicken, chickens like chicken food, and I don't like chicken food."

Festinger's (1962) and Festinger et al.'s (1956) own theory of rationalization, cognitive dissonance, was therefore more specific. It posits that only certain sets of beliefs, desires, and actions can occupy states of consonance or dissonance. Although Festinger did not describe the principle of rational action as such, he grasped it intuitively:[2] People tend to act in a way that maximizes their desires, consistent with their beliefs. Sets of beliefs, desires, and actions that fit this specific principle are consonant; sets that do not are dissonant.[3] Crucially, Festinger also proposed that the state of dissonance is psychologically aversive, motivating people to achieve consonance. If you have already acted (e.g., shooting a man), of course, it is too late to adjust your action. Instead, consonance must be achieved by adjusting your beliefs or desires (e.g., deciding he must have been a threat). This is the essence of rationalization.

Cognitive dissonance is the best-known psychological theory of rationalization; indeed, it is among the best-known psychological theories of anything. Its two main pillars enjoy strong empirical support. First, people change their preferences and beliefs to match actions they have already performed. Specifically, they adopt the beliefs and desires that would have made their past action rational. A classic method that produces such effects, the free choice paradigm (Brehm 1956), is still widely used (e.g., Sharot et al. 2009; 2010; Vinckier et al. 2019). People are given a choice between two things, such as a toaster and a radio, that they value roughly equally. The act of choosing one of these things causes them to value that thing more and the other thing less. This occurs from an early age and in non-human primates (Egan et al. 2007), even when people cannot remember what they chose (Lieberman et al. 2001), and even when the experimenter tricks them into believing they chose something they didn't (Sharot et al. 2010). People's justifications for such choices can be detailed and elaborate (Johansson et al. 2005). Second, experiments confirm that dissonance is psychologically aversive: People say so (Elliot & Devine 1994), and it is also revealed by convert measures such as affect misattribution (Losch & Cacioppo 1990; Zanna & Cooper 1974), psychophysiology (Losch & Cacioppo 1990; Harmon-Jones et al. 1996), and functional neuroimaging (van Veen et al. 2009). In sum, the theory is well known for a good reason.

Yet, while the theory of cognitive dissonance describes the psychological mechanisms involved in rationalization, it does not offer an ultimate explanation – an answer to the question, "Why did it evolve?" It is akin to a theory that explains why we eat by saying, "We are motivated by hunger." Although true and important, such a theory is incomplete: It does not explain why the hunger motive evolved (i.e., because food provides the raw material for metabolism, and that therefore organisms that possess the hunger motive tend to outcompete those that lack it).

Festinger was merely the first in a long line of theorists to explain rationalization while eliding its adaptive function (Harmon-Jones & Mills 1999): Aronson (1968) proposed that dissonance arises most powerfully when actions are incongruent with a person's *self-concept*; Steele (1988; Steele et al. 1993) proposed that dissonance is aversive because people feel that apparently irrational actions threaten their *self-image* or *self-worth*; Beggan (1992) extended this concept to the objects people possess, even when not freely chosen. There are still other possibilities: Perhaps dissonance minimizes post-decisional regret.[4] Each of these proposals elaborates on Festinger's mechanistic account, but they do not offer an ultimate, adaptive explanation for rationalization.

Indeed, among the classic approaches, the descendants of Heider's balance theory come closest to an adaptive rationale. Balance among beliefs and attitudes can be formalized as a form of logical consistency or constraint satisfaction (Read & Marcus-Newhall 1993; Shultz & Lepper 1999; Thagard 1989). If a person represents that *Socrates is a man*, and that *all men are mortal*, but that *Socrates is immortal*, something must give. A well-designed system will repair such inconsistencies in a manner that makes its representations more accurate (Ackley et al. 1985; see also a review by Gawronski et al. 2018).[5] This idea of network repair or cognitive consistency can be fruitfully applied to networks of interrelated beliefs or desires (Cushman & Paul, in press).

But how could it be applied to classic cases of rationalization, in which a person revises their beliefs and desires to match their own past *action*? Suppose that you desire cookies and believe them to be in the kitchen, but go to the living room. Clearly, one adaptive response to this imbalance is to correct your action: To go to the kitchen. This applies the principle of rational action. But now consider the outcome of rationalization: You either decide you didn't want cookies in the first place or convince yourself that they are actually in the living room. Although this achieves coherence of a kind, it certainly does not improve your desires or beliefs. Rather, it takes take a clear error of reasoning and then multiplies it, infecting thought with a pathology of choice.

Rationalization would indeed be counterproductive if our actions were only produced by sound reasons or outright errors. In this case, dissonance would only arise in cases of error – after all, sound reasoning cannot produce actions that violate principles of rationality. And, if rationalization only applied to errors, then it would propagate those errors from action back to our beliefs and desires – a counterproductive result.

Crucially, however, our behavior is influenced by sources other than reason and error. Whether innately, through habit learning, or through cultural learning, our behavior is influenced by processes that are adaptive and yet non-rational. Because these processes share the same ultimate purpose of reasoning – fitness maximization – the beliefs and desires that support reasoning can be improved by learning from the behavioral influence of non-rational sources.

2.2. Self-perception

Bem's (1967) theory of self-perception provides quite a different explanation for rationalization. It denies the two mechanistic pillars of cognitive dissonance: First, that dissonance is psychologically aversive and, second, that underlying desires or beliefs actually change. Rather, it posits that people only change their *perceived* beliefs and desires.

Bem's original statement of this theory was heavily influenced by behaviorism in two ways. First, he assumed that we have no direct introspective access to the mechanisms that produce our behaviors. Second, he claimed that we receive strong reinforcement from others when we can explain our behavior in terms of mental states, and consequently, we often construct such mentalistic explanations. Desiring mentalistic explanations for our behavior, but lacking introspective access, we resort to our best guess: post hoc rationalization.

Although intended as an alternative to cognitive dissonance, self-perception theory stands in its own right as a powerful statement: People are unaware of the causes of their behavior and often attempt to infer these causes by observing their actions. These basic claims are supported by a wealth of experimental research (Devine 1989; Gazzaniga 1967; Greenwald & Banaji 1995; Haidt 2001; Miller & Buckhout 1962/1973; Neisser 1967/2014; Nisbett & Wilson 1977; Wilson 2004). Bem's (1967) statement of this theory may have been influenced by seminal work on emotion misattribution (Schachter & Singer 1962) and confabulation in split brain patients (Gazzaniga 1967). Nisbett and Wilson (1977) later condensed these varied insights into three core claims: People are often unaware of the causes of their behavior; self-reports of the causes of behavior are generated by folk-causal theories; and, therefore, when they correctly report the cause of their behavior, it is usually the result of inference, not introspection.

In sum, there is undeniably something right about the theory of self-perception. Yet, when aimed at the topic of rationalization, it misses two key empirical marks. First, dissonance induces an aversive psychological state that motivates rationalization. Second – and of greatest importance – rationalization changes people's *actual* beliefs and desires, not just their self-perception. For instance, in the free choice paradigm, objects are not just reported to have higher value after being chosen (or lower after rejection), but they are also actually chosen more often in the future (or less often, after rejected).

Like Festinger, Bem did not address what ultimate adaptive purpose might be fulfilled by cognitive dissonance. Instead, he focused on its proximate psychological motivation: The reinforcement of social partners who demand mentalistic explanations of our behaviors. Neither did he squarely address the issue of whether self-perception is usually accurate or inaccurate. In the grip of behaviorism, Bem may have regarded this as beside the point – he was likely skeptical that our behavior relies on structured mental representations at all. Rather, the essence of Bem's claim was simply that we attribute mental states to ourselves by the same processes that we attribute mental states to others.

The present account builds on important insights of self-perception theories (Bem's, and those that followed it), but with two crucial modifications. First, it posits that the function of self "perception" is not merely to satisfy our own curiosity or that of our peers. It also constructs new beliefs and desires based on information implicit in other adaptive control mechanisms. Thus, "perception" is a misnomer: Rationalization is designed not to accurately infer unconscious mental states, but to construct new ones; it is not a discovery, but a fiction. Second, we can, should, and do actually adjust our beliefs and desires to match this fiction. This is adaptive because reasoning and non-rational processes are ultimately trying to maximize the same goal: biological fitness. In other words, rationalization is a fiction, but a decidedly *useful* one. Mixed right, it can be nourishing to drink your own Kool-Aid.

2.3. Responsibility avoidance and impression management

Finally, some theories posit that rationalization is designed not so much to inform others as to *persuade* them, casting your behavior (or other information) in favorable and possibly deceptive light (Tedeschi et al. 1971; von Hippel & Trivers 2011). When risking blame, for instance, we may profess benign motives or faultless naiveté, even at the expense of the truth.

Mercier and Sperber (2011) go so far as to claim that reasoning itself is principally adapted to the problem of changing others' minds, and therefore they interpret rationalization as an adaptive solution to the problem of winning arguments (see also Haidt 2001; Tetlock 2002). Indeed, on their view, reasoning itself is mostly an instance of rationalization. Its goal is to present information to another person in a manner that compels them, by logic or intuition, to accept your conclusion (see also Haidt 2001).

At first blush, such theories seem to explain only why we *express* rationalization to others, but not why we adjust our own beliefs or desires (Tedeschi et al. 1971). Yet, it is also plausible that "true believers" are better deceivers. In other words, the best way to convince others that you shot an (innocent) man for good reason might be to first convince yourself of his guilt (Trivers 2000; von Hippel & Trivers 2011).

This family of theories likely explains a part of the function of rationalization. If the present account also explains a part, then it is a complementary but largely independent explanation.

3. Rationalization as construction

More than a century of research shows that our behavior is influenced by multiple processes (Dolan & Dayan 2013; Kahneman 2011b; Thorndike 1898). One influence is rational planning: considering the likely outcomes of our behavior according to our beliefs, and then choosing the behavior most likely to maximize our desires. Other influences on our behavior, however, are not organized according to the principle of rational action.

A potential function of rationalization, then, is to construct beliefs and desires that are consistent with the adaptive behaviors generated by non-rational processes, and then to adopt them. In other words, like Jason Bourne, rationalization generates new, useful insights by observing the actions we perform thoughtlessly. Later we will view this through a Bayesian lens, as an inversion of a generative model of rational action (Baker et al. 2009) and thus a variety of "inverse reinforcement learning" (Ng & Russell 2000).

To explore the logic of rationalization in more detail, it helps to focus on three potentially non-rational influences on behavior: instincts (innate influences on behavior), conformity to social norms (socially learned influences on behavior), and habits (reinforced behaviors). Although highly simplified, this taxonomy reveals some important insights about the general structure of rationalization, as well as its specific application in different settings. All three discussions depend on the common assumption that non-rational influences on our behavior are nevertheless adaptive. This is a natural assumption, given that our instincts, habits, and norms are all processes shaped by adaptive forces: biological evolution, reinforcement learning, and cultural evolution, respectively.

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

3.1. Rational action as planning

Before considering how instincts, norms, and habits can improve reasoning, we must have a clearer image of how reasoning itself works. Reasoning, sometimes called planning, chooses actions by expected value maximization (Fig. 1a). A simplified model of planning has three parts. First, there is a mechanism for learning a causal model of the world, one's beliefs. This model allows you to predict what is likely to occur in different situations, depending in part on your own actions. Second, there is a mechanism that assigns intrinsic value to certain outcomes, one's desires (also sometimes described as reinforcement or reward). Third, there is a mechanism that chooses actions by maximizing the satisfaction of your desires, given your beliefs. Our next goal is to understand how a system designed this way could extract useful information from instincts, norms, and habits.

3.2. Instincts

Instincts are innate influences on behavior, designed by natural selection, that bias certain actions to be performed in the presence of certain stimuli.[6] For instance, humans instinctively drink when they are thirsty, flee from threats or fight them, reject likely pathogens, fall in love with other humans, and so on. Many of these examples involve very abstract actions (flee) or stimuli (threat). Instincts need not be low-level or concrete, or grounded in a single, well-defined neural mechanism. Rather, *instinct* often describes a very abstract kind of innate mental organization. Its key property is just that the relationship between the stimuli and the actions is innate and direct. For instance, the perception of a threat may directly bias action toward flight. This is what makes instincts different from planning, which would instead require a computation like "fleeing avoids threats, threats might harm me, and I don't like being harmed."

Because instincts are shaped by natural selection, they tend to increase our biological fitness. Similarly, rational planning is designed to increase biological fitness. This is an important part of why rationalization makes sense: It extracts information from one adaptive system (instinct) and makes it available to another (rational planning). If a person's instinct is incongruent with her beliefs and desires, adjusting those beliefs and desires to match her action may ultimately improve them.

For instance, suppose that a person instinctively recoils from snakes. This instinct is adaptive because many snakes are venomous, but she happens to be unaware of this. Rationalizing her instinct (i.e., attempting to explain her act of recoiling in terms of beliefs and desires), she might adopt the belief that snakes are dangerous. Similarly, she could rationalize her behavior by adopting a general desire to be far from snakes, and this is a useful desire. In either case, the outcomes of her future reasoning are improved.

3.3. Rationalization as a form of rational inference

Even if rationalization could possibly generate true beliefs and useful desires, what guarantees that it would do so typically? The specifics of the snake case are suspiciously convenient; this naïf might have concluded instead that snakes breathe fire, shoot crossbows, or dredge up hurtful memories of adolescence. Such beliefs would explain one's instinctive recoiling, but they are false. Or she might have adopted the desire to avoid all animals or all things that are long and straight. What processes could ensure that the rationalized beliefs and desires are, in fact, useful ones?

The construction of new beliefs and desires should presumably be structured as a form of rational inference. In Bayesian terms, the posterior beliefs about snakes ("snakes bite" vs. "snakes shoot crossbows") should be sensitive not just to the likelihood of an action (recoiling) given a percept (snake) and candidate beliefs (e.g., "snakes bite" vs. "snakes shoot crossbows"), but also the prior probability of the candidate beliefs, including their compatibility with other beliefs (e.g., animals can't use crossbows; many long, straight things are perfectly safe).

Indeed, these pieces of information may be integrated in a rational manner, according to Bayes' rule. This form of inference has been well characterized in models of inverse planning (Baker et al. 2009; Ng & Russell 2000). This brings into focus an important dimension of the claim that *rationalization is rational* – it is not just biologically adaptive, but it may also approximate a well-understood form of rational inference. Importantly, however, the approximation of a rational inference (at Marr's computational level) may be quite cognitively simple (at Marr's algorithmic level). These relationships, between rational inference and the actual mechanisms of rationalization, are discussed more fully in section 4.

3.4. Norms

Human psychology is influenced not just by the biological inheritance of natural selection but also by a vast cultural inheritance. And just as biological natural selection ensures that instincts will typically be adaptive, cultural selection ensures that norms will typically be adaptive (Boyd et al. 2011), although maladaptations may arise in each case. Our cultural inheritance takes many forms: concepts (π), artifacts (knives), beliefs (the earth is round), desires (money), norms (drive right, pass left), and much more. The specific form of norms is often transmitted by social conformity. These operate analogously to instincts: just as instincts are innate biases on action, norm conformity may be defined as a set of socially learned biases on action.

As with instincts, norms may be very abstract. For instance, there are cultural norms of cooperation and fairness that generalize over many diverse features of specific cases. Norms may also be redundant with other kinds of cultural influence. For instance, somebody might comply with the Jewish laws of *kashrut* because (1) they wish to get along with their religious peers, or (2) they believe that God asks this of them, or (3) it just feels like the right thing to do. These are, in fact, independent and redundant elements of cultural learning. According to our restrictive definition of norms, only the third influence is sufficiently direct to count as a norm. The first two – a desire to get along and a belief about God – instead influence her behavior indirectly, by reasoning.

Given this homology between instincts and norms, the very same logic that makes instinct a useful target of rationalization therefore also applies to norms. When a person performs a behavior due to cultural influences, she may often be able to extract useful beliefs and desires by rationalizing her action.

Among the indigenous people of Fiji, for instance, it is taboo to eat certain kinds of seafood when pregnant or nursing (Henrich & Henrich 2010). Most of the taboo seafoods are toxic and pose special risks to fetuses and infants, but the people of Fiji do not have precise knowledge of this – indeed, the taboo extends to several closely related seafoods that are actually harmless. Rather, most mothers avoid eating these foods simply because of the norm. Rationalization, however, might lead a mother to the correct belief that these fish are dangerous. Then,

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

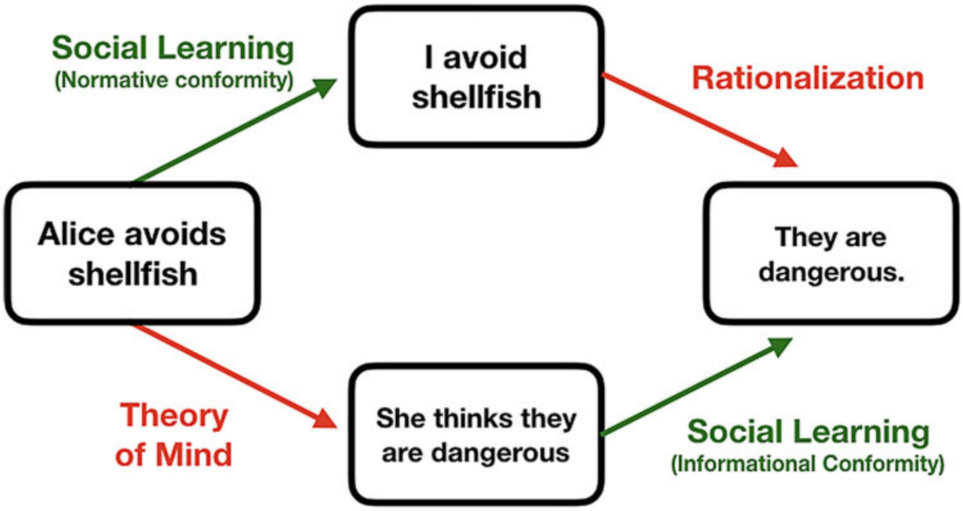

Figure 3. Two parallel pathways of social learning: one characterized by theory of mind and the other characterized by rationalization.

by reasoning, she may generalize other useful conclusions: don't even touch these fish, don't let your baby eat them, don't feed them to sick people, and so forth.

3.5. A "useful fiction": How theory of mind supports social learning

Because norms are commonly transmitted via observation and imitation (Boyd et al. 2011; Cialdini & Goldstein 2004), the benefits of rationalization one's *own* behavior can also be obtained by rationalizing *others'* behaviors. Put more simply, there is a deep homology between rationalization and theory of mind (Fig. 3; see also Bem 1967). This motivates a brief but important detour to consider the relationship between theory of mind, self-perception, and rationalization.

Consider again the seafood taboos of Fiji. A mother might conclude that taboo seafood (suppose it is shellfish) is dangerous to infants by either of two paths: one via rationalization and another via theory of mind. Following the first path, she first complies with the norm itself, avoiding the shellfish simply because others do. Next, observing her own behavior, she rationalizes that shellfish must be dangerous. Following the second pathway, she instead first attempts to understand the behavior of her social partners. This act of mental state inference, or theory of mind, really just amounts to rationalization of others' behavior. She concludes that her social partners must believe that shellfish are dangerous. Then, assuming that they know something she doesn't, she adopts this belief herself.

Each of these paths involves a crucial step in which a belief is extracted from an action. In the first path it is the observer's own action, so we call it rationalization; in the second path it is another person's action, so we call it theory of mind. (This connection, of course, originates with self-perception theory; Bem 1967).

This clarifies two important but distinct functions of theory of mind. Many past treatments assume that theory of mind is designed to infer the *true* causes of another person's behavior (Baron-Cohen 1995; Dennett 1987; Gopnik et al. 1997). Likewise, self-perception theory might be construed as an attempt to accurately infer the true cause of one's own behavior (although Bem himself was agnostic on this point). But we have emphasized an alternative function of rationalization: to *construct* representations that are implied by behavior (one's own, or another person's). This process of construction need not result in a perfectly accurate representation of the causes of behavior in order to be useful (a point made by Dennett [1987] in introducing the intentional stance[7]). To the contrary, rationalization can extract useful beliefs and desires from the influence of non-rational systems – systems whose influence on behavior had nothing to do with those beliefs and desires. Rationalization is, in this sense, a useful fiction. It takes the form of inference, but with a very different function.

This same function can also apply to theory of mind. Theory of mind may often involve useful fictions, in which we ascribe inaccurate causes to others' behavior – goal-directed plans, based on beliefs and desires – even when those behaviors were produced by non-rational processes.[8] Although inaccurate, such ascriptions could still extract useful information for us: true beliefs and adaptive desires.

This perspective has at least one attractive feature: Useful or not, theory of mind seems to involve a great deal of fiction. Despite widespread consensus that human behavior is not exclusively rational, nearly all studies of mental state inference posit a folk theory of *rational* action. Despite more than 40 years of study, there is virtually no research on folk theories of instinct, habit, reflex, and the like – in other words, a theory-of-the-rest-of-our-minds. Moreover, what little research exists suggests that people interpret others' actions as the product of goal-directed reasoning far more than it actually is the cause (Gershman et al. 2016). Similarly, experimental demonstrations of automatic behavior are often surprising to lay audiences, while experimental demonstrations of rational behavior are not. On the useful fiction model, this is because theory of mind is not only designed to infer the true causes of a person's behavior, but also to extract useful beliefs and desires from their behavior even when it was caused by non-rational processes. It is, therefore, biased to perceive all behavior as rational, even though much behavior is not.

3.6. Habits

Habits are a third major non-rational influence on behavior. Habits are learned stimulus-response mappings, often reinforced

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

by reward and punishment (reviewed in Dolan & Dayan 2013). For instance, a person might habitually flip the lights on when they walk into a room because it is typically useful (i.e., rewarding). Each time that the behavior is performed and rewarded, habit learning strengthens the mapping from stimulus to response. As a result, executing habitual action requires little cognitive effort. A person does not have to consider desires ("I need light") and beliefs ("switches cause light") to derive the value of performing an action; rather, the behavior is habitized based on its value in the past. But, for the same reason, habitual control can be inflexible. You might habitually switch the lights on even though you are walking into the room of a sleeping baby and want it to be dark. Rational planning, in contrast, can flexibly adjust to new or unusual circumstances.

Despite these differences, there are two key similarities between habit and reasoning: Both involve learning from direct experience, and both are sensitive to the same rewards. These similarities make it challenging to explain how the rationalization of habitual action could provide new information to a system of reasoning. The challenge has two parts: how to improve beliefs and how to improve desires.

3.6.1. How to improve beliefs

Any experience that trains a habit also ought to inform our beliefs, and thus it is not clear why our habits would imply useful beliefs that we would not already represent explicitly. For instance, the experiences that formed your habit of turning on the lights ought to have also taught you that flipping the switch makes the light turn on – the very belief that you need to flip on the lights by reasoning. In this case, there is no extra information for your system of goal-directed reasoning to extract.

This first challenge has a few simple replies. First, a person may simply have forgotten certain facts, and yet nevertheless have retained an adaptive habit. We have all had the experience of being asked for our opinion on something – a restaurant, a book, a colleague – and being able to recall the valence of our feelings ("I know I liked it"; "something about him gave me the heebie-jeebies") without being able to recall what was eaten, read, or spoken. Even after every detail of an experience evaporates from memory, the residue of our attitudes may remain. According to contemporary theories, this residue – the cached values of objects, events, or actions – is the basis of habits (Dolan & Dayan 2013). By rationalizing habits, we can reconstruct the details that most likely explain them.

Second, and relatedly, a person may have failed to ever formulate the relevant belief (e.g., because they were not paying attention) and yet still have acquired an adaptive habit. Just as it is familiar to have forgotten why we loved a movie or distrusted a person, it is equally familiar to never have been quite sure in the first place. This is possible because habits and beliefs are learned by distinct and dissociable processes (Foerde & Shohamy 2011; Foerde et al. 2013; Knowlton et al. 1996).

3.6.2. How to improve desires

The second and more profound challenge is to explain how adaptive *desires* could also be extracted by rationalization. This challenge is harder because both systems – habit and reasoning – are assumed to begin with the very same set of basic desires: "rewards" (Sutton & Barto 1998) or "primary reinforcers" (Kelleher & Gollub 1962), as they are often called. What information, then, could the habitual system encode that would not already be encoded by the reasoning system? Consider a person who habitually eats cake. Rationalizing this behavior, he concludes, "I like cake." Although true, isn't this information redundant? He ought to have discovered that he liked cake back when he took his first bite.

The answer to this challenge depends on a key insight regarding the nature of value-guided learning and decision making. Often our behavior is organized sequentially, with early instrumental actions chosen because they eventually bring us to intrinsically rewarding states of affairs (Bellman 1954). To plow, sow, and harvest are instrumentally valuable actions, for instance, because they ultimately bring a rewarding feast. A major challenge, then, is to discover or estimate the instrumental value of various actions. This challenge is especially obvious in games like chess: We are attempting to learn the instrumental value of moves (or sequences of moves), which is defined by their probability of ultimately attaining checkmate.

Habit and reason estimate value in different ways. Habit learning involves a backward-looking assignment of value: We wait until checkmate is achieved and then reinforce the sequence of moves that brought us there (Bayer & Glimcher 2005; Glimcher 2011; Morris et al. 2006; Roesch et al. 2007). In contrast, reasoning involves a forward-looking assignment of value: We mentally simulate hypothetical future sequences of moves, attempting to divine whether they are likely to achieve our goal (Dolan & Dayan 2013; Sutton & Barto 1998).

A further benefit of rationalizing habits that depends upon the hierarchical nature of human planning (Badre & Nee 2017; Botvinick 2008; Norman & Shallice 1986). For instance, if our goal is to make coffee, we plan by calling to mind a series of subgoals (grind beans, get filter, heat water, etc.), which may themselves contain subgoals (turn on the faucet, turn on the kettle, etc.). The essential properties of these subgoals are that they are instrumentally valuable given the superordinate goal, and also that they support generalization across diverse circumstances. But discovering this form of instrumental value does not come for free – indeed, a major challenge for current theories of hierarchical planning is to explain how we discover the appropriate ways to carve a task into subgoals (Botvinick 2008; Botvinick & Weinstein 2014; Sutton et al. 1999).

A crucial function of rationalizing habits, then, may be to translate the instrumental value representations of the habitual system into goal (Keramati et al. 2016) or subgoal (Cushman & Morris 2015) representations useful to the goal-directed system. For instance, if a tennis player habitually rushes to volley at the net after serving, this likely reflects the instrumental value of serve-and-volley for winning a point. When rationalizing this behavior, she may say, "my goal was to gain an advantage over my opponent while he was on his heels, in order to quickly win the point." If she internalizes this subgoal, what has she gained? Not a change to the value of winning the point (which both systems represented) or a change to the cached value of serve-and-volley (which the habitual system represented), but a novel subgoal representation ("try to serve-and-volley!") that can improve future planning.

In sum, because values are hard to accurately estimate, and because habit and reason estimate value in different ways, the desires implicated by habitual action may improve our ability to maximize reward by reasoning.

3.7. Hybrid control

Although it is convenient to act as if certain actions are wholly under habitual control, others wholly instinctual, and so on,

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

this is a caricature. Even the simplest targets of rationalization studied in the laboratory – the choice of a toaster over a radio, for instance – are not the product of pure instinct, habit, or norm compliance. Rather, they involve at least some degree of conscious, deliberative planning ("let's see, what could I do with a new toaster?"). More generally, it is disputed whether systems of habit, instinct, or norm adherence could be cleanly severed from reasoning at all (Dayan 2012; Graybiel 2008; Kool et al. 2018).

Yet, while reasoning often contributes to choice, it rarely operates alone. Rather, most behavior is the result of some form of approximate planning – an elaborate background of automatic and non-rational processes that construct a restricted and tractable decision space in which limited rational planning can effectively guide behavior (Cushman & Morris 2015; Dayan 2012; Gigerenzer & Selten 2002; Huys et al. 2015; Keramati et al. 2016; Kool et al. 2018). Instinct guides our minds away from rationally deliberating the possibility of marrying our siblings (Lieberman et al. 2007); habit guides our minds away from the possibility of making coffee by putting bread in the toaster (Morris & Cushman 2017); norms guide our minds away from the possibility of catching a ride to the airport by stealing a car (Phillips & Cushman 2017). Non-rational processes also structure tractable planning by identifying valuable end states (Keramati et al. 2016) or goals (Cushman & Morris 2015). A person may seek revenge instinctually, and yet plot his revenge by reasoning; he may seek cocaine habitually, and yet plan to get cocaine by reasoning; he may seek to divide his resources fairly due to blind norm adherence, but then reason carefully about how the fairest division could be accomplished.

Thus, even when our behavior is jointly determined by the influence of rational and non-rational processes, there is an opportunity for rationalization to extract useful information from the influence of non-rational processes and translate these into a form useful to the rational system.

3.8. Summary: Rationalization is rational

Instincts, norms, and habits shape our behavior in adaptive ways, but not by rational planning based on beliefs or desires. Still, these influences are adaptive: We instinctively recoil from precipices because they *are* dangerous; norm-based food taboos reflect *real* toxins, and habitually flipping a light switch is *usually* a good idea. Rationalization, then, is a useful fiction: When we observe our own behavior, we infer the beliefs and desires that would have been most likely to have caused that behavior, as if it had been an exclusive product of reasoning. Then we adopt those beliefs and desires. This is adaptive because, on average, the new beliefs are true and the new desires promote fitness. For the same reason, theory of mind might often entail a useful fiction as well: By assuming that others' behaviors are rational (when they are merely adaptive), we can extract useful information. Crucially, whether rationalizing our own action or that of others, the process of inferring information from behavior is structured as a rational inference (Baker et al. 2009).

In sum, rationalization exchanges representations of *Do this!* for representations of the type *Believe this!* or *Desire that!* This is a kind of representational exchange: It extracts information implicit in the representations of non-rational systems and transforms it into the format useful to the rational system. The final section expands this view of representational exchange, showing how rationalization is one example of a much broader class of cognitive operations that facilitate the flow of information among distinct systems of behavioral control.

4. A theory of representational exchange

Rationalization extracts information from non-rational systems and makes it available to reasoning. It is apparent how this could improve reasoning, but why would it improve the overall welfare of the organism? In the end, what matters to an organism is not to have true beliefs and useful desires, but to perform the right actions. Insofar as our actions are already appropriately guided by non-rational forces (habits, instincts, and norms), what extra advantage do we gain by improving beliefs and desires?

Properly addressing this question leads to a theoretical framework that encompasses far more than rationalization. Rationalization is just one variety of *representational exchange*: the process of translating information from one psychological system, or representational format, into another. And representational exchange is useful for the whole organism because it organizes information in useful ways – ones that best meet its demands when the information is required. For instance, some ways of representing information demand little computation but are relatively inflexible, getting it right in only a restricted range of cases. Others require greater computational demands but are more flexible, getting it right in a wider range of cases. Representational exchange allows an organism to transform representations of one kind into representations of another, making thought more efficient by balancing the demands of computational effort and flexibility. This more general perspective, a theory of representational exchange, unifies rationalization with many other cognitive operations.

4.1. The structure and function of representational exchange

During rationalization, information flows from non-rational systems to rational ones. Could information flow in the opposite direction – from reason to other adaptive systems, or among the other systems themselves (Fig. 4)? Several examples come to mind. During habitization, choices that were effortful (i.e., rationally planned) become automatic (i.e., habitual). During norm internalization, actions that we observed others perform shape our intrinsic preferences. Although traditionally these processes have been considered unrelated, they are all forms of representational exchange: the sharing of information between distinct mechanisms of behavioral control.

Representational exchange is useful because distinct mechanisms of behavioral control have different ways of representing information and guiding action, each with unique advantages and disadvantages. For instance, habits enable rapid, computationally frugal decision making that is occasionally suboptimal, whereas planning attains greater optimality at the expense of time and effort. Representational exchange allows us to keep thought efficient – that is, to attain the most important opportunities for flexibility and generalization, subject to the resource constraint of a limited cognitive capacity. In this manner it fosters "resource-rational" cognition (Griffiths et al. 2015), improving the overall welfare of the organism.

Consider a simple example. For most people, computing 26 + 52 takes moment of thought, while 25 + 25 comes easily to mind. This reflects two different cognitive organizations. One system encodes a procedure for addition and requires effort to derive specific sums (e.g., 26 and 52). Another system encodes a

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

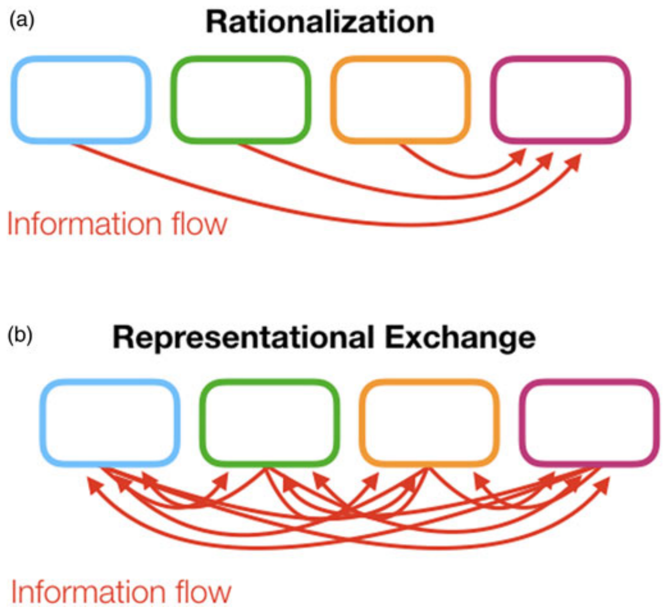

Figure 4. (a) Rationalization is one example of the more general process of (b) representational exchange.

precompiled set of sums – roughly, a table in which one looks up the entry "25 + 25" and retrieves "50." The first requires computation; the second merely requires retrieval.

Why do we have two such systems, and why are certain sums represented one way and other sums another? On the one hand, knowing the rules of addition is useful because it compresses an infinitely large mapping of inputs to outputs (i.e., arbitrary sets of numbers to their sums) via a compact rule. This requires far less memory than, for instance, storing a table of precomputed sums. Although effort is required to compute each sum, this is a worthwhile trade-off as compared with the storage demands of the tabular representation and the learning demands of acquiring it.

On the other hand, certain sums must be computed far more often than others. If you are a cashier who makes change every day, then you do store at least a small table of common sums: "nickel + 2 dimes = quarter," "4 quarters = dollar," and so on. These sums are required so frequently that it would be inefficient to compute them anew each time. Instead, it is worth storing a small cache of common sums for ready and quick retrieval.

Any resource-rational cognitive system must find efficient ways to represent information, managing the competing demands of computational effort, memory, accuracy, and flexibility (Batchelder & Alexander 2012; Griffiths et al. 2015). We must choose whether to represent procedures or merely their outputs (Gershman et al. 2014; Sutton 1991). We must choose when to represent specifics and when, instead, to fall back on generalities (O'Donnell 2015). We must choose when to be exact and when to satisfied with an approximation (Daw et al. 2005; Gigerenzer & Selten 2002; Kahneman 2011b; Kool et al. 2017).

It must be rare that we have attained the optimal balance at any given time – and just as rare that the optimal balance could ever be permanent. Rather, as we learn and change – and as our circumstances and the world around us change – there is a continual demand to adjust the format of the representations that guide our action. This requires mechanisms for representational exchange.

Some forms of representational exchange will pack information into compressed forms, storing outputs, abstractions, and heuristics in place of procedures, specifics, and computations. Other will perform the reverse operation, unpacking information by inferring the more detailed and precise information implicit in outputs, abstractions, and rules. Viewed from this perspective, rationalization is a particular kind of unpacking that occurs in the specific context of choice behavior: Specifically, it unpacks behavior into beliefs and desires. It belongs to a broader family of cognitive operations that facilitate representational exchange – not just from non-rational systems to rational ones, but also among the many systems that contribute to decision making.

Representational exchange can be situated within a broader taxonomy of operations demanded by a successful multisystem cognitive architecture:

1. **Control.** What are the several mechanisms that guide our behavior? What representations and computations do they rely on, and what are the distinctive advantages and disadvantages of each? Much prior research addresses these questions (reviewed in Dolan & Dayan 2013; Kahneman 2011b; Sloman 1996; Squire 2004).
2. **Metacontrol.** Which system, or weighted combination, guides our behavior at any given time? In other words, from moment to moment, how do we decide whether to act habitually, rationally, by instinct, or another way? A growing body of contemporary research addresses this question (e.g., Kool et al. 2017; Daw et al. 2005; Griffiths et al. 2015; Shenhav et al. 2013).
3. **Exchange.** What mechanisms enable the exchange of information between systems (Gershman et al. 2014; Lombrozo 2017)? For instance, how can a behavior formerly produced by reasoning become habitual, and how can a behavior that was formerly habitual influence subsequent reasoning? This is our present focus.
4. **Exchange control.** How do we decide what, and when, to exchange? Assuming that representational exchange can be beneficial in the long run, but also carries immediate costs, how is cost-benefit analysis performed? And, when systems embody conflicting information, which system gets prioritized? These are important issues for further development, but they are not pursued here.

The next few sections present an account of representational exchange somewhat more formally, drawing connections to current computational models of decision making used in psychology, neuroscience, and computer science. Although the main focus is on representational exchange among decision-making systems, it is clear that the concept applies beyond the domain of decision making, and some examples are noted at the end of this article.

4.2. The purpose of representational exchange

The purpose of representational exchange is to make an organism more biologically fit by making its decision making more efficient. Efficiency is a balance of accuracy and effort. Thus, sometimes we increase efficiency by making more accurate decisions; other times, by making decisions faster or with fewer cognitive resources. Efficiency can be optimized by sharing information across decision-making systems in order to give an individual an array of options: more controlled and accurate thought, or more rapid and automatic thought, depending on the circumstances.

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

To describe representational exchange in more detail it is useful to use some formal concepts and notations. These should highlight useful themes for those who are already familiar with them, but without frustrating those who are not. The purpose is not to offer a formal model of representational exchange, which is well beyond the scope of this article. Rather, it is to establish points of contact with formal models of control and metacontrol developed elsewhere.

We envision an organism's life as a kind of Markov decision process. This means that the organism experiences certain *states* of the world, and in each of these states, it performs some *actions*. These actions help to determine the next states it experiences, all of which influence its biological fitness. Any individual's mind can thus be characterized by the probabilistic mapping from states to actions, or *policy*. Colloquially, a policy says: "Here is the thing to do in any given situation" (or "the several things you might do and their associated probabilities"). From the standpoint of natural selection, some optimal policy exists that maximizes expected biological fitness. Nobody actually has an optimal policy, but it is a useful ideal to consider: the total set of instructions for life that maximize your chances of biologically fit children. The closer an organism gets to this ideal, the more fit it is.

As a simplifying assumption, suppose that instinct, norm compliance, habit, and planning each dictate their own specific policy to an organism. In other words, instincts would provide you with one set of instructions; habits with another set of instructions, and so on. These are different mechanisms of behavioral control. In a perfect world every one of these policies would be identical; specifically, they would all encode the optimal policy. In reality, however, different systems are likely to do better or worse in different cases – that is, to recommend more or less fitness-maximizing actions in different states. *Metacontrol* is the problem of deciding how to allocate control to one policy or another in any given situation, or how to blend them.

The goal of representational exchange is to improve the individual policies of each system by transferring information between them. This can improve the overall efficiency of decision making by allowing optimal-but-effortful thought when appropriate, and suboptimal-but-easy thought when appropriate.

4.3. Advantages and disadvantages of control mechanisms

Before asking how these different influences on our behavior might exchange information, greater precision is required on two points: the different formats in which information is represented, and the relative advantages and disadvantages of each format. Briefly addressing these issues will put us in a better position to understand how and why information might be exchanged between representational formats.

4.3.1. Instinct

Instincts are innate mappings from states to actions that emerge regularly in typical development. The advantages of instinct are speed and reliability: They only depend on the development of the organism and not on learning or reasoning, which both take time and are contingent upon unreliable experience. If an organism innately possessed a set of instincts comprising the optimal policy, it would have no need for learning. In reality, however, instincts will not encode the optimal policy because the world changes too fast for biological natural selection to keep pace. Other systems of behavioral control (norms, habits, and reasoning) are useful precisely because they allow an organism to adjust its policy toward fitness maximization more rapidly – on the timescale of a single organism's life, rather than a multigenerational one.

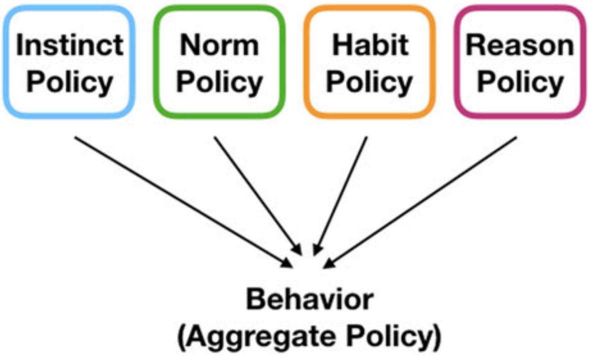

Figure 5. An idealized model of behavioral control.

4.3.2. Instrumental learning

Whereas instincts implement an innately encoded policy, an individual using instrumental learning instead learns a policy by attempting to maximize innately specified rewards. Because instrumental learning occurs within an individual's lifetime, it can improve the individual's policy faster than natural selection. For instrumental learning to improve an organism's fitness, natural selection must assign *reward* to states or actions that reliably increase fitness: things like consuming food, acquiring resources, reproducing, and the like. This often occurs by estimating the instrumental *value* of certain actions – that is, their expected long-run rewards. Current theories of instrumental control tend to use one of two basic ways of estimating value: habit or planning.

Habit. Habit learning is often modeled as a method of estimating the value of every action in every state based on its history of reinforcement (Daw & Doya 2006; Sutton & Barto 1998). The major advantage of habit learning is its low computational demand. First, it only bothers to estimate the value of states and actions that it has actually experienced; for many tasks, this means that the vast majority of conceivable states and actions are ignored. Second, it precompiles (or caches) the instrumental value of actions at the time they are performed, and then draws upon this cached value representation when making future decisions. (This is akin to caching the solution to 25 + 25).

Planning. Planning, like habit, is a variety of instrumental control (Daw & Dayan 2014). It estimates the value of actions prospectively, according to the magnitude and probability of reward of their likely outcomes. When this involves searching over a large model of the potential outcomes it is computationally demanding. Deriving value estimates from an internal model has the advantage, however, of making planning flexible. It can simulate the outcomes of actions it has never performed, it can update its value estimates based on new information, and it can also update them based on new specifications of reward. This may be useful when the agent is tasked with planning toward a specific goal, for instance because of a hierarchical task decomposition (Botvinick 2008; Botvinick & Weinstein 2014; Cushman & Morris 2015; Sutton et al. 1999) or due to social coordination such as joint intentionality (Ho et al. 2016; Kleiman-Weiner et al. 2016; Tomasello 2014).

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

4.3.3. Norms and social learning

Norms are influences on behavior that are learned from others.[9] An extensive literature shows when and why social learning is valuable (reviewed in Richerson & Boyd 2008). The basic premise is simple enough: Because other people are designed to maximize fitness and the things that improve their fitness will often also improve yours, you can improve your own fitness by copying others.

4.4. Varieties of representational exchange

Having reviewed the representational format of several different influences on our behavior and the advantages and disadvantages of each, we can now consider several mechanisms of representational exchange in greater detail.

4.4.1. Rationalization as inverse reinforcement learning

We have already seen that rationalization translates observed actions into useful beliefs and desires. Our next goal is to redescribe this idea both more formally and more abstractly, revealing useful connections to several literatures.

Natural selection and instrumental learning share a common structure: Both are trying to maximize some objective (fitness or reward) by shaping the actions we take in the environments we encounter. An interesting property of rationalization is that it can use the common notion of *objective* to turn one kind of objective (fitness) into another (reward). To see this more clearly, we will begin by representing both processes (natural selection and instrumental learning) identically, as a function:

$$f\,(\underbrace{objective, environment}_{inputs}) = \underbrace{policy}_{output}$$

Later it will be useful to consider the differences between natural selection and instrumental learning. First, however, having defined this function, consider what happens if we flip its direction, swapping inputs and outputs while preserving their mappings. This is the *inverse* function:

$$f^{-1}(\underbrace{policy}_{input}) = \underbrace{objective, environment}_{outputs}$$

This inverse function describes rationalization in very abstract terms: you *input* a sample from the policy (i.e., a set of actions in a state), and you *output* information about the environment and the objective that the agent is trying to maximize. In more ordinary terms, the inverse function could observe a person's behavior (actions) and, on this basis, draw inferences about how the world is (beliefs), and what is valuable (desires). In short, it acts like Jason Bourne.

This basic idea has been widely explored in computer science, where it is often called inverse reinforcement learning (IRL) or inverse optimal control (Ng & Russell 2000). Whereas a typical problem that artificial intelligence (AI) is designed to solve is *choosing actions given an objective* (i.e., reinforcement learning or optimal control), in some settings it is desirable to solve the inverse problem: inferring an objective from a set of observed actions. This goal often arises in social (or multiagent) settings. For instance, if a programmer wishes for a machine to learn by observing humans or to predict human behavior, then the programmer might design the machine to try infer the set of rewards that best explains the human's behavior. Once a set of likely rewards has been observed, the AI can copy human performance by maximizing those rewards itself, or it can predict human behavior by computing which actions would be reward-maximizing for the human.

IRL is easier said than done. In practice, it is often accomplished by some approximation of Bayesian inference. To see how this works, note that we can consider reinforcement learning itself as a probabilistic generative model. That is, given some specification of reward and an environment (composed of many states, actions, and the transition probabilities between them), reinforcement learning algorithms generate a probability distribution over actions – that is, the policy:

$$P(action|reward, environment, state)$$

What IRL seeks, however, is the opposite: the probability of different rewards and environments given an observed action in a given state. This can be computed by inverting the generative model according to Bayes' rule (for brevity, action, state, reward, and environment are now represented by their first letters):

$$P(r, e \mid a, s) \propto P(a \mid s, r, e)P(r, e \mid s)$$

The leftmost term states what we want: inferences about rewards and environments generated by the observation of what a person does, *a*, in some state, *s*. This is proportional to two things we have: the principle of rational action, which derives a policy by reinforcement learning (i.e., $P(a \mid s, r, e)$), and a prior distribution over rewards and environments. Thus, we can guess how the world is, and what is valuable, by inferring the beliefs and desires that would render observed actions rational given our theory of mind.

Notably, a variety of the same Bayesian inversion is essential to computational models of mental state inference (Baker et al. 2009). This could be formalized in the language of a Markov decision process (states, actions, rewards, etc.), but it is more natural to formalize it in the ordinary language of folk psychology (beliefs and desires). The homology between these formalizations is, however, apparent. We begin with a generative model that predicts action on the basis of an agent's beliefs (i.e., its perception of its current state as well as general beliefs about its environment) and desires (i.e., rewards):

$$P(action|desires,\ beliefs)$$

This is given by the principle of rational action. Often, however, our goal is to infer unknown beliefs and desires by observing actions. In this case, we may invert the generative model according to Bayes' rule (again, variables are represented by their first letters):

$$P(d, b \mid a)\ \propto P(a \mid d,\ b)\ P(d,\ b)$$

Thus, we derive a guess about somebody's beliefs and desires given the actions we have seen them perform, $P(d, b \mid a)$, from capacity to predict their actions based on beliefs and desires by the principle of rational action, $P(a \mid d, b)$, and a prior distribution over beliefs and desires $P(d, b)$. These examples illustrate that IRL and theory of mind are, in essence, the same.

What happens if we extend the logic of these computations to a setting where there are multiple forms of behavioral control: not just reasoning, but also instinct, habit, norms, and so forth?

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Crucially, the basic machinery of IRL can work even when an observer assumes a different cognitive architecture than the actor is actually employing. For instance, the actor could be operating with an innate and unchanging policy derived from some process of natural selection (i.e., instincts), in which case its objective is to maximize fitness; nevertheless, the observer could attempt to infer an objective function in terms of rewards stated within the reinforcement learning framework (as well as the structure of their common environment). This could be a useful fiction if the observer is designed as a reinforcement learning agent, and if the fitness objective of the actor is relevant to the reinforcement learning problem. Thus, we shall now stop referring to fitness and desires in common terms as objectives and instead represent the crucial difference between them: One is a property of the world, and another is a mental state. This divide is real, and yet it is bridged by the act of rationalization.

Consider, for instance, an organism rationalizing an instinct. Natural selection has shaped our instinctual policy not according to beliefs and desires, but according to actual facts about the world and biological fitness:

$$f(\underbrace{fitness, environment}_{properties\ of\ world}) = policy$$

where natural selection defines the objective in terms of fitness. Yet, during rationalization, the inverse function computed is:

$$f^{-1}(policy) = \underbrace{desires, beliefs}_{mental\ representations}$$

Put in plain words, whereas fitness and actual environmental conditions shaped our instincts, rationalization extracts *desires/rewards* (a mental representation of an objective function that roughly correspond to the objective of fitness) and *beliefs* (a mental representation of the environment that roughly correspond to actual environmental conditions). Inferences about the causes of our actions become a bridge that translates properties of the world into mental representations of those properties. This makes sense because, roughly speaking, the ultimate function of belief is to represent true properties of the world, and the ultimate function of desire is to represent the fitness consequences of these properties.

In sum, rationalization approximates a form of rational inference and thus can be understood as a variety of IRL at Marr's computational level – its function is to extract useful information from observed actions. This does not imply, however, that rationalization always involves Bayesian inference at a mechanistic level. In some cases, it may, but in other cases relatively simple cognitive processes, akin to those identified by Heider and Festinger, may approximate the rational inferences described above.

As we shall see next, a benefit of construing the present theory of rationalization in terms of these more formal concepts, and at Marr's computational level, is that it makes apparent the relationship between rationalization other forms of representational exchange.

4.4.2. Habitization: Cached value and cached policy

During rationalization, information is extracted to improve reasoning; we next consider several ways in which analogous processes can extract information to improve habitual action.

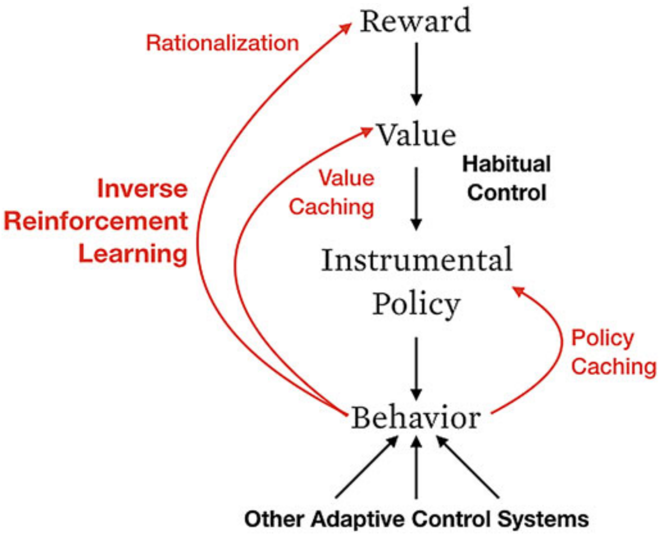

Figure 6. The representational hierarchy of instrumental learning and decision making. Information can be stored at any level: Cached policy, cached value, or beliefs and rewards.

According to several theories, habits can be understood as cached representations of instrumental value – that is, the expected value of actions, in terms of long-run reward. Plausibly, then, useful habits can be constructed by extracting information about value from the actions selected by other systems. Here, again, we envision this as a process of IRL, but this time the goal is to derive a *value function* that can be cached for habitual action:

$$P(v \mid a, s) \propto P(a \mid s, v)P(v \mid s)$$

Thus, while certain instincts, or norms, may not themselves depend on any representation of value, still we may update our cached value representations (i.e., habits) by inferring the values that are consistent with the actions performed.

In contrast to such value-based models of habit, however, some alternative theories posit that habits depend upon cached *policy* – direct stimulus/response mappings, with no representation of value. On this view, habits are chunks of policy "stamped in" through mere repetition (Dezfouli & Balleine 2012; 2013; Dezfouli et al. 2014). For instance, by repeatedly tying our shoes in a particular manner, a specific sequence of actions is chunked into an easily retrieved bundle. Such cached policy representations introduce a new target for representational exchange. When a person's action is determined by instinct, planning, or even value-based habitual action, this action may be cached directly as a policy weight – its informational content thus exchanged into a new format particularly suited to efficient online execution.

Thus, while value- and policy-based theories of habits have been viewed as competitors, they may instead be viewed as complementary representations within a unified scheme (Fig. 6). The hallmark of an instrumental system (whether habit or planning) is that rewards shape values, and values shape policy. Planning makes maximal demands on online computation – it must derive value representations from basic representations of reward and the environment, and then derive a policy from those values. Value-based habitual control (value caching) requires substantially less online computation – it can derive a policy from a cached value

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

representation. Value-free habitual control (policy caching) makes the minimal demands – it simply enacts the stored policy. Thus, value- and policy-based theories of habit need not be considered as rivals, but as distinct points on a common spectrum.

4.4.3. Offline planning as representational exchange

Reasoning is used not only to choose current actions, but also to support offline planning – that is, simulating actions, anticipating their consequences, and then caching the resulting values for rapid decision making in the future (Buckner & Carroll 2007; Davidson et al. 2009; Daw et al. 2011; Gershman et al. 2014; 2017). This is sometimes described as the rational system training the habitual system. It was introduced to the reinforcement learning literature as the Dyna architecture (Sutton 1991), and it is spontaneously deployed by humans (Gershman et al. 2014).

Imagine, for instance, a downhill skier competing in the Olympics. She is given a few opportunities to walk the length of the course, building a mental model of it. She then returns to her hotel room and repeatedly visualizes the process of skiing the course. During this offline simulation, she is able to precompile a habitual policy. As a result, when she actually traverses the course at speeds approaching 80 miles per hour, she can quickly execute her policy without online planning.

Recently, there has been some interest in interpreting imagination, hypothetical and counterfactual thinking, and even causal judgment as forms of offline planning (e.g., Gershman et al. 2017; Icard et al. 2018; Lombrozo 2017; Morris et al. 2018). The perspective offered here makes apparent the connections between these traditional areas of psychological research and machine learning methods such as Dyna and Monte Carlo tree search (Browne et al. 2012), which similarly involve offline model-based evaluation to improve a cached value or policy representation.

4.4.4. Representational exchange during social learning

During social learning, information is exchanged between individuals. Social learning need not also involve representational exchange, but it often does. In fact, across several diverse literatures on social learning, culture, and norms, one of the most prominent themes is that social learners extract many different kinds of information when observing others. Representational exchange occurs when the kind of representations guiding an actor's behavior are not the kind extracted by an observer.

In comparative and developmental psychology, observational social learning is often organized into two broad types: imitation and emulation (Tomasello et al. 1987; Whiten et al. 2009). Imitation occurs when a learner directly copies the overt behavior of a social target. It could be thought of as something like policy caching: a direct update to the probability of performing certain actions in certain states. In contrast, emulation occurs when a learner infers the goal behind somebody's behavior and then adopts only the goal. This allows the learner to design his own policy to attain that goal, potentially by different means. Thus, it involves a form of IRL, or theory of mind, in which the behavioral policy of another individual is used to generate and then adopt new representations of reward or value.

Social psychologists have developed related taxonomies. Many authors have noted that the effects of norm learning can be deeper or shallower (Cialdini & Trost 1998; Deutsch & Gerard 1955; Kelman 1958). These can be construed as different varieties of representational exchange (Morris & Cushman 2017). For instance, sometimes people follow a norm because they represent it as something that other people do and care about – so, they comply with the norm in order to get along. This is sometimes called compliance or normative conformity. The norm is represented explicitly and can be thought of as a part of a person's world model; it influences behavior via planning about the likely consequences of compliance versus non-compliance.

Other times people follow a norm because they believe that the behavior of others tells them something true and important about the world – for example, "If everyone is avoiding the roast beef, there must be something wrong with it." This is often called informational conformity. It corresponds to a variety of IRL in which inferences about others' beliefs become the basis for updating your own. It could be thought of as a deeper form of norm compliance because it gives a person a reason to comply with a norm even in the absence of an audience.

Finally, sometimes people internalize a norm – that is, they come to directly value whatever it is that the norm prescribes. This might be because norm compliance becomes habitual (i.e., its value is cached), or more deeply still, because they represent the very acts implied by the norm as intrinsically rewarding. This final possibility is the most permanent and influential because reward representations are less subject to subsequent update than value representations, and because they exert an influence on both goal-directed and habitual control systems.

In summary, imitation, emulation, compliance, informational conformity, and internalization all embody different models of how an organism can update its representations in response to the same social observations (Morris & Cushman 2017). From the perspective of representational exchange, a major goal of the organism is to update the specific representations that will render the social information most useful to its future behavior.

4.4.5. Other forms of social learning

Another form of social learning is instruction, which is, roughly, one person telling something to another person. This involves an exchange of information between individuals, but it may still maintain the representational format from teacher to learner. Thus, for instance, a teacher might convey her beliefs to a learner, who would then update her own beliefs; or, the transfer might occur from value to value, reward to reward, or policy to policy. Like mere imitation, these cases involve the exchange of information between individuals, but not an exchange of representational formats.

Learning by instruction – that is, people talking to each other – may be a very important setting for rationalization of a different type, however. Mercier and Sperber (2011) propose that we often rationalize our behavior to other people through explicit verbal communication (e.g., argumentation) in order to attempt to influence their beliefs and desires in ways that are useful to us. This illustrates the way in which the current theory of rationalization and other past theories may explain distinct and complementary aspects of the phenomenon.

A final form of social learning is evaluative feedback, in which a teacher provides rewards and punishments to a learner in order to exploit their capacity for instrumental learning to ultimately shape their policy. This interesting and complex form of representational exchange is, however, beyond our present scope (but see Ho et al. 2017).

4.5. Beyond decision making: Other forms of representational exchange

Representational exchange is useful not just for decision-making processes, but in many other areas of cognition as well. The examples of computing "25 + 25" versus "26 + 52," for instance, do not really belong to the same general category of decision making as, say, planning a trip to the grocery store, tying one's shoes, or leaping away from a snake. Still, we have relatively computationally cheap, precompiled knowledge of common sums (such as 25 + 25) and relatively more computationally intensive methods of deriving uncommon sums (such as 26 + 52). And there are circumstances in which it will be optimal to exchange information between these formats.

Representational exchange has been well explored in at least one domain that isn't principally about decision making: thought experiments and other forms of imaginative learning (Lombrozo 2017). These are cases in which an individual has some kind of intuitive knowledge of a phenomenon (e.g., the behavior of a physical system) and uses this intuitive knowledge as a basis for improving their explicit theory of that phenomenon (e.g., a new theory of physics, such as gravity or relativity). This may appear to be a form of alchemy, conjuring gold from iron filings. In reality, it is more akin to rationalization. First, a generative process is trained by experience; for instance, the visual system might learn to anticipate the motion of physical bodies under various conditions. Next, an individual inspects the information that the generative process makes *explicit* (e.g., an intuition about how physical objects will interact) and uses that information to draw rational inferences about information that is merely implicit (e.g., the laws of motion). Just as rationalization extracts structured information from precompiled value representations, a thought experiment about physics extracts structured information from precompiled predictive perceptual representations.

The resulting information – a theory of physics, for instance – is far more computationally expensive to use under many conditions. (In other words, it is harder to determine how an apple will fall from a tree by applying Newton's laws than by relying on precompiled predictive perceptual representations). But this explicit theory is also far more flexible, allowing us to solve problems for which we have no adequate precompiled predictive perceptual representations such as landing an astronaut on the moon.

Although rationalization, theory of mind, and thought experiments might ordinarily be considered very distinct phenomena, they share both structural and functional similarities – an unpacking of implicit information compressed into a narrow format. Noticing these similarities may help us to develop an abstract framework for understanding how, why, and when people engage in representational exchanges of any kind (see Batchelder & Alexander 2012).

Thought experiments are common in philosophy, and they are sometimes used in the process of achieving reflective equilibrium. This is a canonical case of representational exchange. When seeking reflective equilibrium, one contrasts intuitive judgments about particular cases with principled rules or reasons that govern those cases, and then seeks the minimal modifications to both that bring them into alignment (Daniels 2003). It is commonly used during moral reasoning: Individuals seek to bring their intuitive judgments of particular cases into alignment with a more general normative theory, and this involves revision to both the particular judgments and to the theory itself.

Reflective equilibrium is attractive to philosophers who want to take intuition seriously without giving it absolute priority. They emphasize that intuition is the result of adaptive processes, which could include habit learning, natural selection, and cultural evolution, among others (e.g., Railton 2014). This echoes a basic argument offered here: The policy recommendations of non-rational systems can be used to improve reasoning, at least from the standpoint of fitness maximization.

Yet, skeptics have reasonably countered that even if intuition tends toward adaptive outcomes, it is at best a heuristic approximation (e.g., Greene 2014). Why, then, ought intuition ever be favored over reasoning? If a person has devoted appropriate mental effort to reasoning, then what *superior* policy recommendation could arise from the heuristic representations of non-rational systems?

One simple reply is that a person's beliefs might be wrong. But this is a shallow response; presumably, we are often quite confident in our reasoning, and yet it still conflicts with our intuitions. A deeper reply is that rationalizing non-rational adaptive processes can improve intrinsic rewards maximized by reasoning. By itself, reasoning has no method for questioning or improving its own goals; it is, in Hume's (1739) words, "a slave to the passions." Even a system of reasoning possessed of perfect causal knowledge and unconstrained by computational resource might be improved by information from intuition. Of course, *improvement* in this case means more biologically fit; whether this is the kind of improvement that philosophers want is a question for philosophers.

This highlights the relationship between rationalization and the naturalistic fallacy (that is, deriving what "ought to be" from what "is"). In a specific setting – adopting new desires for instrumental learning from observations of one's own or others' actions – we have considered an adaptive rationale for this inference. Of course, it is not always the case that people do the best (i.e., most fit) thing; on average, however, there is something important to be learned about what one *ought* to do simply by observing the actions that people, including yourself, actually perform.

4.6. Summary

More than a century of psychological research shows that our behavior is a product of multiple systems. Given that these each have relative advantages and disadvantages, there is good reason to exchange information across representational formats, gradually optimizing the manner in which information is represented. This perspective highlights the common function of many different processes that convert information between diverse representational formats and across individuals (Fig. 7).

A theory of representational exchange suggests several important topics for further development. First, assuming that representational exchange has a computational cost, how is it allocated efficiently? In other words, is there a mechanism of exchange control that regulates when and where representational exchanges of different types occur? Second, when information conflicts arise between systems during representational exchange, how are they adjudicated? One obvious prediction is that an individual will not adopt new beliefs or desires when the action produced by a non-rational system can easily be explained away as a mistake. For instance, if someone reflexively jumps away from a fake rubber snake, they might not adopt the belief that rubber snakes are dangerous. Rather, they might conclude that this was a misfiring of a non-rational system, and discount it as a source of information. It remains to be seen how this inference can be formalized.

5. Conclusion

Why did I do that? We ask ourselves this question often – perhaps more often than we would like. Why do we bother?

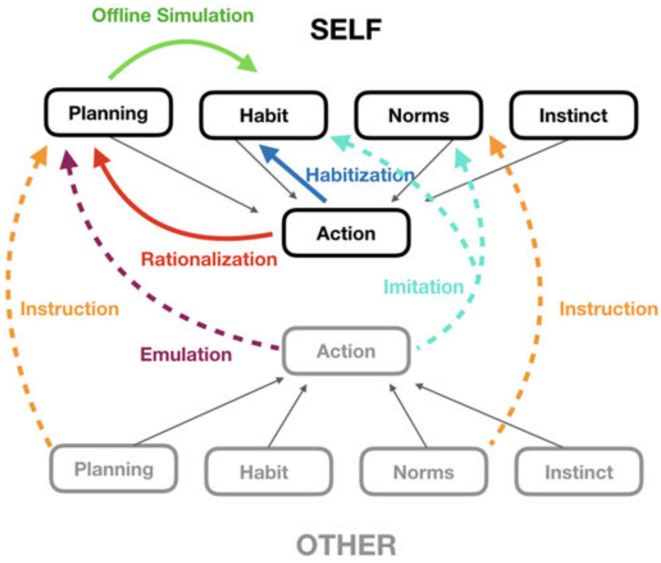

Figure 7. Varieties of representational exchange.

Many past approaches suppose that we are motivated by self-discovery, or the desire to explain ourselves to others (e.g., Bem 1967). Possibly, for instance, our behavior is guided by unconscious reasoning. If so, then *Why did I do that?* is equivalent to *What were my reasons?* – we are trying to divine the obscure beliefs and desires that underwrite our unconscious decision making. Perhaps this is often so.

But human action is also shaped by non-rational forces. In these cases, any answer to the question *Why did I do that?* that invokes belief, desire, and reason is at best a useful fiction. Whether or not we realize it, the question we are actually answering is: *What facts would have made that worth doing?* Like an amnesic government agent, we are trying to divine our programmer's intent – to understand the nature of the world we inhabit and our purpose in it. In these cases, rationalization implements a kind of rational inference. Specifically, we infer an adaptive set of representations that guide subsequent reasoning, based on the behavioral prescriptions of non-rational systems. This inference is valid because reasoning, like non-rational processes, is ultimately designed to maximize biological fitness. It is akin to IRL as well as to Bayesian models of theory of mind, and thus it offers a new interpretation of the function of these processes.

Viewed in this light, rationalization is just one example of a broader set of representational exchange mechanisms. Our minds are built to exchange information across multiple systems of behavioral control, allowing for an efficient balance of computation and flexibility during decision making. By perceiving the common function of these processes, we can better comprehend their structure.

Notes

[1] I am indebted to Daniel Wegner and Joshua Greene for this succinct statement of the mirror-image relationship between rationality and rationalization, borrowed from their lecture slides for the undergraduate course "Social Psychology."

[2] Festinger and subsequent theorists of cognitive dissonance often wrote of actions that "follow from" or do not "follow from" cognitions, and this vague concept appears to have typically captured the principle of rational action – that is, in practice, actions were judged by these theorists to follow from beliefs and desires when those actions maximize desire satisfaction conditioned upon beliefs, relative to alternative actions.

[3] In many cases dissonance could arise among mental states, such as beliefs and desires; he writes, for instance, "A person may think Negros are just as good as whites but would not want any living in his neighborhood." Festinger was primarily interested, however, in the relationship between mental states and actions.

[4] Thanks to an anonymous reviewer for this suggestion.

[5] Although such network repair can be used to improve systems of interrelated beliefs and desires, the relevant processes of update, repair, or dissonance reduction may still be infected by cognitive or attitudinal biases (Gawronski et al. 2018).

[6] According to this restrictive definition not all innate mental structure is an instinct. For instance, we have innate mechanisms of perception, memory, attention, etc., but these are not instincts because are not usefully construed as innate mappings from stimuli to action. Similarly, our innate capacity for rational planning is innate, but it is not an *instinct*. Our innate desires are not instincts either (e.g., the desire not to be thirsty, not to be attacked, not to be sick, and to be loved). These desires bias action, but not directly. Instead, their influence is mediated by reward learning mechanisms like planning and habit. Thus, although we both innately desire not to be thirsty, and also instinctively drink when thirsty, these are distinct and redundant mechanisms.

[7] Dennett's point was that the intentional stance can be a useful fiction for the purposes of describing, explaining, and predicting certain kinds of agents or systems. This is surely true, and the present proposal identifies a further, distinct use of such fictions: To improve one's own reasoning by *adopting* the fictive beliefs and desires.

[8] As Dennett (1987) points out, we might even rationally adopt this intentional stance toward non-agents, such as the process of natural selection.

[9] Thus far we have acted as if the mind contained some distinct, explicit norm-based policy – that is, as a mapping from a state ("going to a friend's house for dinner") to an action ("bring a bottle of wine") that reflects social learning alone. In reality, however, socially learned norms likely influence behavior not through some proprietary representation, but by altering the kinds of representations used by other mechanisms of behavioral control. These include the basic constituents of instrumental learning: representations of value, reward, and the world (i.e., your causal model; see Morris & Cushman 2017). Thus, for instance, the norm of wine-gifting might be represented in terms of the instrumental value of the gift or in terms of a belief that such behavior is expected by social partners. The unifying theme is not the nature of the representation, but the manner in which it was acquired: by social learning.

Acknowledgments. Thanks to Adam Bear, Sam Gershman, Dan Gilbert, Joshua Greene, Mark Ho, Max Kleiman-Weiner, Victor Kumar, Michael Littman, Rebecca Saxe, several anonymous reviewers, and members of the Moral Psychology Research Laboratory for their thoughtful, detailed, and generous feedback. This work was supported by grant N00014-19-1-2025 from the Office of Naval Research and grant 61061 from the John Templeton Foundation.

Open Peer Commentary

Rationalizations primarily serve reputation management, not decision making

Sacha Altay and Hugo Mercier

Institut Jean Nicod, Département d'Études Cognitives, École Normale Supérieure, EHESS, PSL University, CNRS, 75005 Paris, France.
sacha.altay@gmail.com hugo.mercier@gmail.com
https://sites.google.com/view/sacha-altay/home
https://sites.google.com/site/hugomercier/

doi:10.1017/S0140525X19002115, e29

Abstract

We agree with Cushman that rationalizations are the product of biological adaptations, but we disagree about their function. The data available do not show that rationalizations allow us to reason better and make better decisions. The data suggest instead that rationalizations serve reputation management goals, and that they affect our behaviors because we are held accountable by our peers.

Cushman suggests that his individualistic account of rationalization is complementary to a social account of rationalization based on reputation management. However, the two accounts make conflicting predictions, so that data about rationalizations should help tell which account better explains features of rationalizations.

If our ability to rationalize evolved by improving practical decision making, its expression (when we feel the need to rationalize) and content (how we rationalize) should not be strongly modulated by the presence of others, and it should not (as a rule) lead to the acceptance of inaccurate beliefs (since those are unlikely to improve practical decision making). By contrast, if our ability to rationalize evolved by serving reputation management ends, its expression and content should be strongly modulated by the presence of others, and it could lead to the acceptance of inaccurate or practical maladaptive beliefs, in a trade-off between the practical and the social value of the beliefs (see Kurzban & Aktipis 2007; Mercier 2012).

On the first point – when people feel the need to rationalize – we rely on results from the cognitive dissonance literature, given that cognitive dissonance is a driver of the need to rationalize. Experiments have shown that the feeling of dissonance is highly modulated by social context: Participants are less likely to experience dissonance and to rationalize their actions when these actions are private rather than public (e.g., Tice 1992; see also Leary 1995; Tedeschi & Rosenfeld 1981). More generally, people seem more concerned with *appearing* consistent than with *being* consistent (Tedeschi et al. 1971).

Second, the content of rationalizations depends more on what we think others will deem acceptable than on the likelihood of improving practical decision making, as if the rationalizations were generated by an internal "press secretary" (Kurzban & Aktipis 2007; see also Haidt 2001; Nisbett & Wilson 1977). This would explain, for instance, why cognitive dissonance manipulations only affect explicit beliefs (Gawronski & Strack 2004). If the rationalizations that cognitive dissonance gives rise to aimed at improving our practical decision making, they should affect implicit and explicit beliefs, since both matter for practical decision making. By contrast, if these rationalizations serve social goals, only explicit beliefs – the ones we can share – matter. Moreover, the influence rationalizations have on our explicit beliefs can be practically detrimental. In all the standard cognitive dissonance paradigms, dissonance reduction leads to less accurate beliefs – for example, that a truly boring task isn't really that boring (Festinger & Carlsmith 1959).

When rationalizations affect our actions, they often do not do so in a way that is compatible with Cushman's account. Many experiments on reason-based choice (Simonson 1989; for reviews, see Mercier & Sperber 2011; Shafir et al. 1993) have shown (1) that often participants make worse decisions, from a purely practical point of view, when they engage in rationalizations, and (2) that these decisions fit with the expectations participants have of what decision will look best. In the classic reason-based choice study (Simonson 1989), asking participants to justify their decisions (and thus to provide rationalizations) pushed them to make suboptimal decisions (as they were influenced by logically irrelevant factors). However, the participants thought these arguably less practical rational decisions were easier to justify and less likely to be criticized. In another study, Briley et al. (2000) have shown that asking participants in Asia and the United States to justify their decisions affected these decisions, but in opposite ways as a function of the culture to which they belonged. In Hong Kong, participants who provided rationalizations were more likely to choose a compromise option, while participants in the United States were less likely to do so. In each case, the deviations were in line with the society's cultural values, and thus with which rationalizations would be more likely to be accepted. In both studies, rationalizations influenced participants' decisions in a way that made them arguably less practically rational, but more socially acceptable (see also Baumeister 1982; Baumeister & Cairns 1992).

One of the strengths of Cushman's account is that it explains why people's behavior would be guided by past rationalizations: If rationalizations were a purely social tool, why let our future behavior be guided by them? Here, we suggest that Cushman neglects one of the main reasons why rationalizations are socially effective: They commit the speaker to the rationalizations that the speaker offered. By providing rationalizations, we convey the following information: (1) I share the same values as you; (2) I made this decision because of these values; and (3) in the future, I will keep following these values. If (3) were absent, that is, if speakers didn't commit to their rationalizations, the rationalizations would be largely worthless. By being committed to what they say, senders implicitly acknowledge that if their message is found to be unreliable, they will pay reputational costs (e.g., see Mazzarella et al. 2018; Vullioud et al. 2017).

The fact that we are committed (to some extent) to our rationalizations explains why we stick to them: If we didn't, we'd suffer social costs. Rationalizations are thus no different from any other form of commitment. For example, when we morally condemn a behavior, we commit to not engaging in this behavior ourselves, and we dislike people who engage in behaviors they have previously condemned (Jordan et al. 2017). Indeed, our need to appear consistent has long been associated with reputation management (Leary 1995; Sperber 2000; Tetlock 1992).

Last but not least, if our ability to rationalize had evolved to improve practical decision making, the human exception would be quite puzzling from an evolutionary perspective. Being influenced by multiple "non-rational processes" (target article, sect. 1.1, para. 7) and having to make complex decisions are not exclusive to humans. Our social account of rationalization avoids this pitfall: Rationalization is human-specific because the human social niche has exerted selective pressures on our ability to manage our reputation so as to compete in biological markets (Mercier & Sperber 2017).

Means and ends of habitual action

Samantha Berthelette[a] and Christopher Kalbach[b]

[a]Department of Philosophy, University of California, San Diego, La Jolla, CA 92093-0119 and [b]Department of Philosophy, Florida State University, Tallahassee, FL 32306-1500

sberthelette@ucsd.edu ckalbach@fsu.edu

www.smberthelette.com www.ChrisKalbach.com

doi:10.1017/S0140525X19002127, e30

Abstract

Cushman claims that post hoc rationalization of habitual behavior can improve future reasoning. His characterization of habits includes two components: (1) habitual behavior is a non-rational process, and (2) habitual behavior is sometimes rationalized. We argue that Cushman fails to show any habits that are apt targets for rationalization. Thus, it's unclear when – if ever – rationalizing habits would improve reasoning.

Fiery Cushman's account of rationalization involves two main claims: that the beliefs and desires appealed to in rationalizing explanations are constructed post hoc, and that these constructions can help improve future reasoning. While we agree with Cushman regarding the roles that non-rational systems such as social norms (sect. 3.3) and instincts (sect. 3.2) play in rationalization, we believe that he extends his account too far with his discussion of habits (sect. 3.6). Cushman gives two examples to motivate the claim that rationalization extracts implicit information from habitual actions to guide future behavior. However, it seems that these examples are either better explained as intentional actions or they do not involve rationalization at all. In either case, the role of habitual action in Cushman's account of rationalization needs revision.

Post hoc rationalization is effective so long as the agent can construct a belief and desire that favor that particular behavior token. For example, suppose you are on a diet but decide to eat the last cookie anyway, thereby experiencing a lapse in self-control. When asked why you ate the cookie, you are asked to state the beliefs and desires that would thereby make your action rational. In other words, you are asked to give a means-end justification. In response, you might construct a post hoc justification for yourself that makes your action seem rational rather than weak-willed. For example, "it would have gone to waste if I didn't eat it." If not letting food go to waste is a reasonable end, then, *ceteris paribus*, eating the cookie is a justified means to achieve that end. You might even believe that this was the operative reason for your behavior. However, it was your desire to eat the cookie that played the operative role.

Means-end explanations are likewise constructed when we rationalize behavior influenced by social norms and instincts. In each case of rationalization, we construct a belief and desire that favor the behavior in question – thus attributing reasons to non-rational processes. However, it is less clear when and why we would rationalize our habitual actions.

When we act out of habit, the resulting behavior typically either (1) will map onto the right circumstances to produce the intended result or (2) will "misfire." If the habitual action maps onto the right circumstances, then the action accomplishes the very end(s) for which the habit was formed in the first place. In such cases, the habitual behavior token is supported by a belief and desire. But these were not constructed post hoc. Rather, the habitual behavior token is best explained as being caused by a standing belief and desire (see Mele 2007). There is no rationalization of the behavior because the relevant beliefs and desires were there all along.

Now consider Cushman's primary example given to explain the post hoc rationalization of habits: turning on lights in a sleeping baby's room (sect. 3.6). Suppose that I turn on the lights in this scenario out of habit. Suppose further that I developed this habit because it is typically useful to turn lights on when I enter dark rooms. However, in this particular case, it was not useful. Indeed, turning on the lights directly opposes my supposed end: namely, making sure the baby is still asleep. This example thus fits into the second category of habitual action. The habitual action token is a "misfire," which means it did not map onto the right circumstances. I might have some end met by turning on the light (e.g., being able to see clearly), but that end still does not make my behavior rational. It is not rational because my end of keeping the baby asleep is more important to me than being able to see clearly. So, instead of constructing new beliefs or desires that were not operative at the time of turning on the light, the natural response would be to say something like "I forgot the baby was sleeping."

"Misfire" cases like the sleeping baby example leave little room for post hoc rationalizations because they are not supported by a belief and desire. My habitual behavior token did not map onto the right circumstances, and the underlying cause of the behavior was in opposition to my end. Instead of using means-end language, I simply admit my mistake. If rationalizing habits does "provide new information to a system of reasoning" (p. 26), this example does not seem to provide much evidence for that claim. Indeed, it is still unclear what sort of habitual action *would* provide that cognitive benefit.

Let us now turn to Cushman's other example of habitual action. This is the example of the person who habitually eats cake (sect. 3.6.2). If our cake eater does not recognize any salient reasons to stop eating cake, it is unclear what role rationalization could play. However, he might rationalize his behavior if he does recognize salient reasons to stop eating cake. But if he continues to eat cake despite recognizing these reasons, there appears to be more going on here than mere habitual action. In particular, our cake eater seems to be suffering from a lack of self-control. Once *akrasia* is introduced as the cause of his behavior, it is no longer a clear example of behavior generated by non-rational processes. This is because akrasia is generally understood as intentionally doing *x* even though the agent has judged it to be all-things-considered better to do *y* than *x* (e.g., see Davidson 1970). In other words, akrasia is the result of at least some rational processes rather than non-rational ones. If Cushman's goal is to identify scenarios in which "the rationalization of habitual action could provide new information to a system of reasoning" (sect. 3.6, para. 2), then the cake-eating example seems to miss the mark as well.

Perhaps habitual action does play an important role in rationalization, but revision and clarification are needed this aspect of Cushman's account to be useful. As it stands, it is unclear when – if ever – we rationalize habitual behavior in the first place.

Rationalization is a suboptimal defense mechanism associated with clinical and forensic problems

Stuart Brody[a] and Rui Miguel Costa[b]

[a]Department of Psychology, Charles University in Prague, Praha 1, Czech Republic and [b]William James Center for Research, ISPA – Instituto Universitário, 1149-041 Lisboa, Portugal.
stuartbrody@hotmail.com rcosta@ispa.pt
www.DrStuartBrody.com en.ispa.pt/faculty/rui-miguel-costa

doi:10.1017/S0140525X19002073, e31

Abstract

Cushman argues that "rationalization is rational." We show that there is reasonable empirical clinical and forensic psychological evidence to support viewing rationalization as a quite suboptimal defense mechanism. Rationalization has been found to be associated not only with poorer emotional development, but also with a broad range of antisocial behavior, including not only shoplifting, but also pedophilia and murder.

Cushman asserts that rationalization is rational (the non-rational process of rationalization is a useful fiction that improves subsequent reasoning, and thus it pays to rationalize). In this way, rationalization is considered adaptive as it would facilitate the putative "ultimate" goal of reasoning: maximizing biological fitness. However, there is ample evidence that rationalization does not improve reasoning. There is a substantial difference between the word "rational" in the sense used by economists (and to some extent in the target article), which reflects nominal desires and beliefs, and the common usage of the word "rational," which often implies good or wise. However, beyond the circular level of something being termed rational because it seemingly fulfils one's nominal immediate desires, rationalization can be a defense mechanism that has been found to be associated with not only a developmentally lower level of psychological functioning, but also with a broad range of criminal and interpersonally destructive behaviors.

Psychological defense mechanisms are processes that operate largely outside of conscious experience and serve to reduce anxiety at a cost (distortion of reality). They can be conceptualized along a developmental hierarchy, reflecting both the ages at which a defense might be normal and the degree of psychological impairment if used in adulthood. The defense of rationalization allows individuals to cope with emotional conflict or other stressors by creating incorrect explanations that can alleviate their distress. Hence, the perception of reality is distorted in the sense that the correct, but painful, explanations, are avoided (Knoll et al. 2016). Rationalization is generally found to be in the group of immature (or maladaptive) defense mechanisms (Andrews et al. 1993). Rationalization continues to show a decline even in the course of adulthood (Diehl et al. 2014), which speaks not only to a maturity issue, but also possibly learning that it is not a long-term efficacious strategy (it might also reflect greater survival among persons less likely to use rationalization).

Nearly a century ago, Taylor (1923) observed that the process of rationalization is associated with less, rather than more, mental awareness, as a result of suppression of thought. This blocking of thought and reason is part of why rationalization was later found to be a developmentally immature psychological defense mechanism. Far longer ago, an astute observer of human behavior noted: "And oftentimes excusing of a fault Doth make the fault the worse by the excuse" (Shakespeare, *King John.* Act 4, Scene 2).

The defense of rationalization has been found in many studies to be associated with antisocial behavior and antisocial personality disorder (a pervasive predatory or criminal orientation that is resistant to treatment). Various psychiatric diagnostic systems even include rationalization in the criteria for antisocial (sometimes termed dissocial) personality disorder (American Psychiatric Association 2000; World Health Organization 1993). The subgroup of disavowal immature defenses (rationalization, denial, and projection) are especially predictive of antisocial personality disorder (Blais et al. 1996). Perhaps unsurprisingly, rationalization of violence is associated with adolescent antisocial behavior (Calvete 2008), and "high self-esteem" can exacerbate the process of rationalizing antisocial behavior in predisposed children (Menon et al. 2007).

In addition to antisocial behaviors, rationalization might also be a risk factor for the development of posttraumatic stress disorder (Price 2007).

Although Cushman defines rationalization as concocting desire or beliefs after an action so as to make the action appear rational, one prominent theory of criminal behavior asserts that neutralizations (a largely conscious cousin of rationalization) can precede an antisocial act, so as to decrease any possible internal resistance to perpetrating the act (Sykes & Matza 1957). Of course, after the first offense, it becomes more difficult to disentangle neutralizations and rationalization.

Rationalization might contribute to the development of alcohol abuse and alcohol dependence (Wombacher et al. 2019), and rationalization is differentiable from the well-known defense of denial in alcoholics (Ward & Rothaus 1991).

The breadth of antisocial behavioral associations with use of rationalization is noteworthy. Rationalization is associated with engaging in "repeated and unwanted attempts by one person to initiate or maintain an intimate relationship with a specific, targeted, and unwilling other" (Brownhalls et al. 2019). Although such behavior might in itself be seen as consonant with Cushman's view of rationalization being a rational means of improving fitness, rationalization is also associated with perpetration of pedophilia and other child sexual abuse in both sexes (deYoung 1989; Jimenez 2015; Neidigh & Krop 2015; Rush Burkey & ten Bensel 2015).

Rationalization has been found to be highly prevalent among shoplifters (Cromwell & Thurman 2003). Rationalization has also been identified as one of the psychological processes that facilitate being a perpetrator of multiple murders and genocide. This applies in studies of a contract killer (Levi 1981), perpetrators of the Rwandan genocide (Bryant et al. 2018), and Nazi physicians (Lifton 1986). It might well be argued that part of the appeal of Hitler's *Mein Kampf* (Hitler 1925) for the German voters was his rationalization that Germany could not have lost World War I because of its own intrinsic failures, so it must have been due to the Jews. In this case, the immature defense mechanisms of rationalization and projection are intertwined. Some of Hitler's followers likely accrued fitness benefits for a few years, but eventually, Germany lost yet another World War.

In conclusion, rather than the non-rational process of rationalization engendering useful fictions that improve subsequent reasoning, it has been found to be associated with real-world destructive behavior and no sign of improved reasoning even within the destructiveness.

Rationalization and self-sabotage

Jason D'Cruz

Department of Philosophy, University at Albany, State University of New York, Albany, NY 12222.
jdcruz@albany.edu www.jasondcruz.com

doi:10.1017/S0140525X19002231, e32

Abstract

In making the case that "rationalization is rational," Cushman downplays its signature liability: Rationalization exposes a person to the hazard of delusion and self-sabotage. In paradigm cases, rationalization undermines instrumental rationality by introducing inaccuracies into the representational map required for planning and effective agency.

Fiery Cushman's account of the rationality of rationalization is both surprising (because it contradicts the folk wisdom that rationalization is paradigmatically irrational) and unifying (because it offers an evolutionary story for the emergence of rationalization and situates rationalization in the broader theoretical framework of "representational exchange"). However, in highlighting the potential adaptive benefits of being an organism that rationalizes, Cushman downplays rationalization's signature liability: It exposes a person to the hazards of delusion and self-sabotage.

Cushman presupposes that rationalization always happens after action. But the ambit of rationalization is wider. Decisions *not* to act are also rationalized. In addition, rationalization can be *anticipatory*, clearing away hurdles of caution and of conscience. "Pre-violation" rationalization serves to defuse an anticipated threat to the moral self, allowing people to do wrong while feeling righteous (Shalvi et al. 2015). This kind of rationalization is more difficult to square with Cushman's account of the rationality of rationalization, because it puts into focus desires whose satisfaction subjects know or suspect to be self-undermining or morally dubious.

Consider the case of a scientist who refuses to adjust the course of her research program despite the urging of her peers who that worry it is fundamentally unsound. Since it is more pleasant to inhabit the fantasy world where she is a misunderstood genius than the real world where she is an ordinary thinker with a lot of work ahead of her, she may rationalize her intransigence by concocting a story about the inability of her peers to comprehend her profoundly original ideas. Or consider the man who lies to his wife about his ballooning credit card debt. He may rationalize his dishonesty by telling himself that disclosure and transparency would cause his partner unbearable anxiety. These workaday examples conform to the basic pattern of Cushman's account: First, the subject makes a decision or performs an action, then "concocts the beliefs or desires that would have made it rational." (target article, sect. 1.1, para. 1) Although rationalizers don't have unfettered freedom to concoct whatever they want (they must work with the evidence at hand), rationalization is nonetheless an essentially creative endeavor (D'Cruz 2015).

Subjects will sometimes manifest awareness of their concoctions as such. As a result, raising the stakes may induce a person to abandon their rationalizing postures when their most cherished aims are threatened (Gendler 2007, p. 244). A looming tenure case might bring the scientist to take her colleagues' criticism more seriously; the specter of a separation might bring the husband to appreciate the moral weight of duplicity.

However, as Cushman notes, in many other cases "people don't just tell a story, they actually make themselves believe it" (sect. 1.1, para. 4) In such cases, rationalizers' concoctions get added to their stock of beliefs. Ramsey (1931) famously characterized belief as a map by which we steer. The liability of rationalization is that inaccuracy introduced into the map undermines a person's ability to plan intelligently, or as Cushman puts it, "to discover or estimate the instrumental value of various actions" (target article, sect. 3.6.2, para. 2) The deluded scientist may stick to her guns even when her career is threatened; the deluded husband may fail to reckon with his moral failings even when his cherished relationship hangs in the balance. These individuals are self-sabotaging not simply because they fall short of some putatively objective standard of theoretical rationality; they are irrational in the basic sense that their actions fail to realize the satisfaction of their own desires and the realization of their own goals. As Cushman puts it, in these cases rationalization "takes take a clear error of reasoning and then multiplies it, infecting thought with a pathology of choice" (sect. 2.1, para. 7)

Cushman explicitly concedes that rationalization "will be occasionally be maladaptive." (3) However, he insists that "on average, over time, it pays" (sect 1.1, para. 10). The postulate that on average rationalization "pays" is surely an empirical conjecture. The vindication of such a conjecture depends on the extent to which rationalization plays the "information exchange" role, extracting largely accurate information from adaptive but non-rational processes, and the extent to which rationalization subverts goal-directed activity by constructing a wishful map of the world in lieu of an accurate one. A further possibility is that rationalization is irrational (insofar as it subverts agency) yet adaptive (insofar as it improves fitness).

We might taxonomize rationalization into (1) information-extraction rationalization, which takes (adaptive) instinct, conformity to norms, and habit as its input, and (2) reality-distorting rationalization, which takes wishful thinking as its input. We would then want to know how these two types of rationalization are related to each other and whether they are realized by the same mental processes. We would also want to know whether an individual's susceptibility to the reality-distorting species of rationalization is accompanied by a heightened capacity for the information-extraction species, and also whether efforts to avoid reality-distorting rationalization have costs in terms of the fitness advantage conferred by information exchange. Clear thinking and moral integrity require that we be vigilant about debunking stories that serve to justify bad decisions and actions. (Honest friends play a central role in this.) How are we to think about the value of such vigilance in the light of Cushman's central thesis?

Rationalization is rare, reasoning is pervasive

Audun Dahl and Talia Waltzer

Psychology Department, University of California, Santa Cruz, CA 95064.
dahl@ucsc.edu https://esil.ucsc.edu/people/audun-dahl/
twaltzer@ucsc.edu https://sites.google.com/site/taliawaltzer/

doi:10.1017/S0140525X19002140, e33

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Abstract

If rationalization were ubiquitous, it would undermine a fundamental premise of human discourse. A review of key evidence indicates that rationalization is rare and confined to choices among comparable options. In contrast, reasoning is pervasive in human decision making. Within the constraints of reasoning, rationalization may operate in ambiguous situations. Studying these processes requires careful definitions and operationalizations.

Much is at stake in debates about rationalization. To rationalize is to take one's non-rational action and "[concoct] the beliefs or desires that would have made it rational" (target article, sect. 1.1, para. 1). If rationalization were widespread, we would have to abandon a premise of much discourse: that our beliefs and actions are generally based on reasons. Imagine a person arguing for climate policies by stating: "I believe humans have caused global warming, but my belief is not based on reasons." The absurdity of the statement exemplifies how human discourse usually assumes that rationalization is a curious aberration, not our default mode of operation. In support of this assumption, we propose that rationalization is rare and reasoning is pervasive.

Cushman's account of rationalization begins with the claim that people "rationalize all the time" (sect. 1.1, para. 2) and "are never fully certain of what to believe and what to value" (sect. 1, para. 3). We term this the *ubiquity hypothesis*. Cushman supports the ubiquity hypothesis with two kinds of evidence: negative evidence that many behaviors cannot be based on reasoning and positive evidence that many behaviors are caused by something other than reasoning. (We assume that Cushman views his own arguments as reasons that support his theory, not as rationalizations.)

We contend that the reviewed research offers little evidence for rationalization or its ubiquity. As negative evidence of rationalization, Cushman discusses a study of vacation preferences (Sharot et al. 2010). Each participant rated how they would feel, from unhappy (=1) to extremely happy (=6), about various vacation destinations. Next, researchers picked pairs of destinations that the participant had rated identically and told the participant, falsely, that they had chosen one destination over another, for instance, Thailand over Greece. On average, participants who were told that they had chosen Thailand increased their post-choice rating of Thailand by 0.08 points (Sharot et al. 2010, Fig. 1). Cushman describes the participants' responses as "gross errors" and "rationalization" (sect. 1.1, para. 4). However, ratings of, say, 5.00 and 5.08 on a 6-point scale represent near-identical attitudes (Krosnick 1999): Either rating means that participants were very happy to travel. Insofar as the ratings are practically indistinguishable, choosing one over the other would not be an error, let alone a gross error. Moreover, the study did not document that participants concocted beliefs or desires to explain their shifts in ratings, as would have been required by the definition of rationalization.

Indeed, non-rational processes have mostly been found to influence choices among near-equivalent options (Krueger & Funder 2004; Turiel 2010). For instance, incidental disgust can make ratings of others' actions slightly more negative, say from 3 to 14 on a 100-point scale (Wheatley & Haidt 2005). In contrast, incidental disgust rarely, if ever, causes categorical shifts in judgments (e.g., from "okay" to "wrong," Landy & Goodwin 2015; Pizarro et al. 2011). For more consequential choices, for instance, about saving instead of taking a life or vacationing in York instead of New York, evidence for non-rational influences is scant (Dahl et al. 2018; Royzman & Hagan 2017).

We find the positive evidence for rationalization similarly unconvincing. Cushman considers three "non-rational" influences on behavior: norms, innate instincts, and learned habits (for critiques of innate vs. learned behaviors, see Dahl 2019; Spencer et al. 2009). On Cushman's account, people often follow norms because "it just feels like the right thing to do" (sect. 3.4, para. 2). Discussing a study on food avoidance among pregnant and nursing women in Fiji (Henrich & Henrich 2010), Cushman writes that the women avoid certain foods "simply because of the norm" and are unaware that the foods "pose special risks to fetuses and infants" (sect. 3.4, para. 4). However, interviews with the mothers contradict this interpretation: Most believed that the foods *did* pose risks to their fetuses and infants, and many had additional reasons for avoiding the foods (e.g., social disapproval; Henrich & Henrich 2010, supplementary materials). The study did not show that people followed norms merely because it *felt* right – rather, participants expressed reasons for their actions.

Psychological research shows that reasoning plays a pervasive role in human decisions (Dahl et al. 2018). By reasoning, we mean the formation of beliefs or decisions in accordance with considerations that people articulate and endorse (Dahl & Killen 2018). Reasoning is evident in many areas of psychology, operating quickly or slowly and often in concert with emotions (Adler & Rips 2008; Ajzen & Fishbein 2005; Harman 1986). Our work has shown that children and adults provide reasons that largely explain their social and moral judgments (see Dahl & Killen 2018; Turiel & Dahl 2019). In recent research on moral judgments, we interviewed adolescents and adults about pushing one person to his death in order to save five others (Dahl et al. 2018). In Study 1, 81% said it would be wrong to push the one person; nearly of half of them reasoned that the victim had no involvement in the situation. Study 2 manipulated victim involvement by stipulating that the person to be pushed had tried to kill the five persons; now, only 27% thought pushing the person was wrong. As in numerous studies, participants provided reasons for their judgments that, in experimental manipulations, proved to guide their judgments.

The abundant evidence for reasoning, and the limited evidence for rationalization, lead to our conclusion that reasoning is pervasive and rationalization is rare. This conclusion is both reassuring and forward-looking: reassuring because it supports the assumption that our beliefs and actions are largely guided by reasons, and forward-looking because it points to new areas of research on rationalization. We expect that rationalization is particularly common in ambiguous or challenging situations in which individuals struggle to decide among comparable options (Dahl & Waltzer 2018; Turiel & Dahl 2019). Studying rationalization in ambiguous situations is a valuable topic of inquiry, as long as researchers recognize how many of our everyday situations are far less ambiguous (Dahl 2017). More research is needed to identify the contexts in which rationalization occurs, and to separate reasoning and rationalization through careful definitions and operationalizations.

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Rational rationalization and System 2

Wim De Neys

LaPsyDE, UMR CNRS 8240 and Université de Paris, Sorbonne, 75005 Paris, France.
wim.de-neys@parisdescartes.fr www.wdeneys.org

doi:10.1017/S0140525X19002048, e34

Abstract

In this commentary, I highlight the relevance of Cushman's target article for the popular dual-process framework of thinking. I point to the problematic characterization of rationalization in traditional dual-process models and suggest that in line with recent advances, Cushman's rational rationalization account offers a way out of the rationalization paradox.

The dual-process framework (e.g., Evans & Stanovich 2013; Kahneman 2011a) has long conceived human thinking as an interplay of fast and intuitive processing ("System 1" thinking) and slower, more demanding deliberate processing ("System 2" thinking). The characterization of rationalization in this popular framework is often problematic.

On one hand, rationalization is typically conceived as epiphenomenal. It is considered as a mere "making-up-excuses-after-the-facts" in which reasoners post hoc look for justifications for an (often inappropriate) intuitively cued choice. For example, one might think here of classic reasoning and heuristics-and-biases tasks in which people fail to correct biasing intuitions but are afterwards all too eager to find reasons to support their (erroneous) intuition (e.g., Evans & Wason 1976; Kahneman 2011a). As such, rationalization would have no functional "rational" role to play in sound reasoning.

At the same time, dual-process theorists also tend to characterize rationalization as a deliberate (System 2) process. Indeed, people often spend considerable time and effort to come up with justifications and rationalize their answers (Pennycook et al. 2015; Wason & Evans 1975). This poses a puzzle. Why would we waste scarce resources on a pointless epiphenomenon? The fact that the human cognitive miser – who typically prefers to refrain from demanding deliberation – nevertheless engages in it to rationalize its behavior suggests that rationalization must serve an important function (Evans 2019). Unfortunately, this functional role of rationalization has received little attention in traditional dual-process models.

However, recent dual-process work has started to hint at a possible role in social communication and argumentation (Bago & De Neys 2019; De Neys 2017; Evans 2019). A key observation is that reasoners rationalize not only incorrect intuitions but also correct ones. One intriguing finding comes from two-response studies in which reasoners first have to answer as fast as possible with the first response that comes to mind and afterwards can take the time to deliberate and give a final answer (Bago & De Neys 2019; Newman et al. 2017). Results indicate that sound reasoners do not necessarily need to deliberate to correct an initial erroneous intuition (e.g., "10 cents" in the infamous bat-and-ball problem); their initial intuitive response is often already correct. However, without subsequent deliberation, they struggle to give an explicit justification of their (correct) intuitive answer (Bago & De Neys 2019). In other words, good reasoners seem to intuitively know the correct response, but don't seem to know why it is correct in the absence of further deliberation. This indicates that sound reasoners do not necessarily deliberate to correct their intuition but to rationalize it and look for an explicit justification.

As Mercier and Sperber (2017) have stressed, such a justification process in which we look for explicit reasons in support of our intuitions can be critical to efficiently sway others. Clearly, if I want to convince my peers that my solution to a problem is right, I will be more successful when giving them an explicit, verifiable argument than by simply telling them that I "felt" it was right (Bago & De Neys 2019).

Whereas the recent dual-process findings (and the work of Mercier & Sperber 2017) point to a possible functional role of rationalization in social persuasion, Cushman's account points to an additional contribution to our own "internal" information processing. In my opinion, such "internal" and "external" functions do not need to be mutually exclusive. However, my goal here is not to comment on the specifics of Cushman's proposal. The key point I want to highlight is that by pinpointing a rational role of rationalization, Cushman's work offers dual-process theorists a possible way out of the rationalization paradox.

As I tried to clarify, the lack of a functional account of rationalization is problematic for dual-process theories. If rationalization is not rational, it would be hard to explain why we spend our dearest resources on it and still survive as a species. Cushman builds a convincing case for the rationality of rationalization. Therefore, any dual-process proponent (or critic) should take note of it. My hope is that this will instigate renewed empirical research on rationalization in the dual-process field.

Rationalization in the pejorative sense: Cushman's account overlooks the scope and costs of rationalization

Jonathan Ellis[a] and Eric Schwitzgebel[b]

[a]Department of Philosophy, University of California at Santa Cruz, Santa Cruz, CA 95064 and [b]Department of Philosophy, University of California at Riverside, Riverside, CA 92521-0201.
jellis@ucsc.edu jonathanellis.ucsc.edu/
eschwitz@ucr.edu faculty.ucr.edu/~eschwitz

doi:10.1017/S0140525X19002152, e35

Abstract

According to Cushman, rationalization occurs when a person has performed an action and then concocts beliefs and desires that would have made it rational. We argue that this isn't the paradigmatic form of rationalization. Consequently, Cushman's explanation of the function and usefulness of rationalization is less broad-reaching than he intends. Cushman's account also obscures some of rationalization's pernicious consequences.

According to Cushman, rationalization occurs when a person has performed an action and then concocts beliefs and desires that would have made the action rational. He argues that the function of rationalization is to transfer information among the processes that influence our behavior. We argue that Cushman-style rationalization is only one form of rationalization, not the paradigmatic form. Consequently, Cushman's explanation of the function and usefulness of rationalization is less broad-reaching than he intends. Cushman's account also obscures some of rationalization's pernicious consequences.

In one of the earliest psychological treatments of rationalization, Ernest Jones wrote:

> Everyone feels that as a rational creature he must be able to give a connected, logical, and continuous account of himself, his conduct, and opinions, and all his mental processes are unconsciously manipulated and revised to that end. (Jones 1908)

We rationalize not only our conduct or actions, but also our opinions or judgments. And we rationalize our actions not only after we perform them, but also before we perform them and sometimes as a condition of performing them.

At the newsstand, the cashier accidentally hands Dana $20 in change instead of $1. Dana notices the error and wonders whether to point it out. She thinks to herself, "What a fool! If he can't hand out correct change, he shouldn't be selling newspapers. And anyway, last week he sold me a damp newspaper, so this turnabout is fair." Consequently, Dana keeps the $20. Despite these thoughts, if Dana had seen someone else receive incorrect change in a similar situation, she would have thought it plainly wrong for the person to keep it.

What is rationalized in this case is both a judgment (that keeping the change is morally fine) and a behavior before it occurs (keeping the extra change), which the judgment is used to license. Dana's reasoning is epistemically flawed in the way characteristic of many rationalizations: It is distorted by an irrelevant factor (financial self-interest) that is not acknowledged. Furthermore, since the act of rationalization precedes Dana's action of walking away with the $20, it is possible that had Dana been unable to concoct a minimally adequate rationalization, she would not have performed that action (Kunda 1990).

Dana's type of rationalization is a – maybe *the* – paradigmatic form of rationalization, the kind of rationalization frequently lamented and colorfully depicted in literature, philosophy, and psychology. People usually conceive of rationalization pejoratively. Not only does Dana's type of rationalization involve an epistemically distorting factor, but it often licenses selfish, immoral, or harmful actions.

We call this kind of rationalization *rationalization in the pejorative sense*. Rationalization of this kind occurs when a person favors a particular conclusion as a result of some factor (such as self-interest) that is of little justificatory relevance. The person then seeks an adequate justification of that conclusion, but the very factor responsible for their preferring that conclusion distorts this search for justification. As a result of an epistemically flawed investigation, the person endorses a justification that makes no mention of the distorting factor guiding their search (Schwitzgebel & Ellis 2017). Human beings rationalize in this way all the time – about climate change, tax cuts, morality, spirituality, relationships, nearly everything of importance.

Unlike rationalization as Cushman characterizes it, this kind of rationalization targets not just behavior, but also judgments, and it can as easily license prospective behavior as justify past behavior. Also, rationalization in this sense is epistemically flawed in a way Cushman's characterization doesn't capture.

We have two concerns. First, although Cushman identifies one form of rationalization and a possible function of that form of rationalization, it is not obvious how Cushman's functional explanation, which appeals to informational transfer, would generalize to paradigmatic forms of rationalization like Dana's. If the account does not generalize, then Cushman's attempt to fill the lacuna he sees in the literature is limited.

Possibly, Cushman would not extend his evolutionary hypothesis to other forms of rationalization. He might argue that he has identified the primary, most fundamental, or evolutionarily or developmentally earliest form of rationalization, and that cases like Dana's are derivative. More needs to be said.

Second, Cushman's characterization of rationalization and his emphasis on its usefulness obscure rationalization's frequent and serious consequences. Paradigmatic cases of rationalization in the pejorative sense involve biased, motivated reasoning that is epistemically flawed and often recruited to justify immoral or harmful actions. And when people rationalize as Dana does, they impair the social evaluation of reasons. In a social exchange of reasons, you might defend conclusion A by appeal to reason B. Ideally, in what we call "open exchange," B is your real reason for concluding A: Not only do you believe that B supports A, but this belief is also the primary cause sustaining your belief in A. Your interlocutor thus has three ways to change your mind: Either show that A is false or unsupported, show that B is false or unsupported, or show that B does not support A. In cases of rationalization like Dana's, B isn't the real basis of belief in A, and if B is shown false or unsupportive of A, Dana will (if sufficiently motivated) just reach for a new reason C. The real basis of Dana's belief remains hidden; it's not really open for peer examination.

We all know how frustrating it is to argue with someone about politics or about why they (not we) should perform such-and-such unpleasant duty, when they are rationalizing. The person offers reasons for their we-think-mistaken view; we undercut their reasons; they simply shift to new reasons – possibly multiple times. Their stated reasons aren't their real reasons. Rationalization in the pejorative sense is the psychological process behind this phenomenon. Rationalization prevents open dialogue on which socially embedded cognition crucially depends. It is one of the most fundamental epistemic vices.

There is some usefulness to rationalization. There may even be substantial epistemic benefits, as Cushman argues. (See also Bortolotti 2015.) But rationalization has a bad name for good reason, and that will be missed on Cushman's theory.

Belief as a non-epistemic adaptive benefit

Rebekah Gelpi, William Andrew Cunningham and Daphna Buchsbaum

Department of Psychology, University of Toronto, Toronto, ON, Canada M5S 3G3.
rebekah.gelpi@mail.utoronto.ca individual.utoronto.ca/rgelpi/
cunningham@psych.utoronto.ca http://scslabuoft.wordpress.com
buchsbaum@psych.utoronto.ca http://www.cocodevlab.com

doi:10.1017/S0140525X19002206, e36

Abstract

Although rationalization about one's own beliefs and actions can improve an individual's future decisions, beliefs can provide other benefits unrelated to their epistemic truth value, such as group cohesion and identity. A model of resource-rational cognition that accounts for these benefits may explain unexpected and seemingly irrational thought patterns, such as belief polarization.

Rationalization is often conceived of as a betrayal of epistemic truth: Someone who rationalizes believes that the reasoning behind their actions or thoughts can be explained as the result of a rational process, but their inability to access the true motivations behind their behavior leads them to draw a mistaken conclusion. In this conceptualization, a rationalizing actor is doubly irrational: Not only are their actions not governed by reasoning, but they have also concocted an imaginary, if plausible, narrative that recasts them as rational.

We propose that beliefs can serve several functions, only one of which is representing epistemic truth. Cushman describes rationalization as eliciting a "useful fiction," which already gestures at a process in which an individual's representation of the world is not entirely faithful, although it is still useful for the purpose of improving one's future decisions and beliefs with respect to these representations. But these useful fictions need not be in the service of these ultimate goals; for example, shared belief is also an important element to social cohesion and group identities (Echterhoff et al. 2009; Jost et al. 2008). Sharing a belief with those in one's community is therefore beneficial not only when (and because) that belief is true, but also when (and because) it provides an individual with the benefits of a group, such as a sense of belonging and easily accessible shared knowledge.

This belonging is not elicited by social conformity alone. Indeed, the "shared reality" generated by a community relies on its members' certainty that they believe in it on the merits of the evidence (Echterhoff et al. 2009); in other words, they are rationalizing about why they hold these beliefs. In turn, the rationalization that results in a group's system of shared belief and belonging is powerful. As a result, to reap the benefits of belonging, it is advantageous, and in fact rational, to ignore evidence that would require believing something that threatens one's relationship to a social group. This phenomenon is especially apparent with ideological and moral beliefs, as well as other beliefs that can become central to one's identity (Kruger & Dunning 1999; Jost et al. 2003), and these beliefs can become quite resistant to change, with members of a community dismissing contradicting evidence and even experiencing altered memory and perceptual judgment rather than giving up on a shared belief (Van Bavel & Pereira 2018).

Trusting the beliefs shared by one's social group – and "outsourcing" one's own cognition to depend on knowledge held by others in their community – can also reduce the need to engage in cognitively effortful reasoning about a variety of daily needs, even those as basic as one's source of food or shelter (Sloman & Rabb 2016). Similar to the heuristics and biases encountered in perceptual judgment, the reliance on shared community beliefs reflects a need to optimize one's limited resources for individual cognition and reasoning. In a variety of cognitive tasks and situations, the manifestation of biases such as anchoring may reflect the rational use of these resources, accounting for the costs of additional computation against the diminishing improvements in outcome they provide (Lieder et al. 2018). The use of comparatively cheap heuristics may predispose humans to systemic biases in certain cases, but the cost of these biases is outweighed by the benefits of saving limited cognitive resources.

Existing "resource-rational" approaches to modeling cognition have generally treated beliefs as valuable to the extent that they represent the world accurately. However, by accounting for the utility that beliefs can provide unrelated to their truth value, such as in providing members of a community a sense of group identity and belonging, or in their ability to bolster the effectiveness or usefulness of other elements of one's belief system, we can better understand the mechanisms that motivate people to process information in a biased fashion and fail to update their beliefs as a result. This could allow us to clarify how phenomena that seem to defy traditional conceptualizations of rational belief, such as belief polarization – the strengthening of opposing views in two different individuals or groups after observing the same data – may be understood as the result of a resource-rational process.

The phenomenon of belief polarization leads to a calcification of increasingly extreme views that become progressively more resistant to change (Lord et al. 1979). The motivation behind failing to integrate information that goes against one's existing beliefs, or even in fortifying one's existing views against this information, appears on the surface to be irrational. However, the same biases that allow people to rely on their local majorities as a source of shared beliefs that can offer better social outcomes are those that can predispose them to be especially motivated to maintain these beliefs, even when they are incorrect or lead to conflict. This motivation is further fortified because giving up on certain beliefs may threaten one's broader worldview or the safety of one's position within a social group; keeping certain fictions may be preferable if they improve the function of one's causal understanding of the world.

The adaptive value of beliefs, and the rationalizations that bring them about, goes beyond simply improving an individual's predictions and decisions. Beliefs can also be a formative component of an individual's self- or group identity, to the point where it may sometimes be more rational, given limited cognitive resources, to dismiss evidence that would threaten them than to adjust one's views to account for new data. In an increasingly polarized social climate, in which people seek out information that confirms their own views and reject data that do not fit with their a priori model of the world, fully understanding the non-epistemic motivations for maintaining beliefs in the face of negative evidence is critical to developing methods to challenge the entrenched, unquestioned thought patterns that belief polarization gives rise to.

Ideology, shared moral narratives, and the dark side of collective rationalization

Jesse Graham

Eccles School of Business, University of Utah, Salt Lake City, UT 84112.
jesse.graham@eccles.utah.edu
https://eccles.utah.edu/team/jesse-graham/

doi:10.1017/S0140525X19002267, e37

Abstract

This commentary extends the target article's useful concepts to consider collective instances of representational exchange. When groups collectively rationalize their actions, entire networks of beliefs and desires can be created and maintained in the form of shared moral narratives and system-justifying ideologies. These collective rationalization cases illustrate how adaptive advantages can come at the expense of the truth.

Cushman portrays representational exchange as a set of exclusively within-individual processes, even when based on one's observation of another's action (e.g., Fig. 7 in the target article). But processes like rationalization often occur at the collective levels of groups, societies, and cultures. Collective rationalization – whereby a group's collective action leads them to update or create new shared beliefs and desires – may be rational as well. Just as individual rationalization can extract useful information from non-rational sources like instincts and habits, communal processes can cohere moral intuitions and norms into the shared "useful fictions" of shared moral narratives.

As the target article makes clear, even individual rationalization is inherently social. When individuals rationalize, they are extracting useful information from social inputs. We need to update our beliefs and desires to better align with those around us. We do this via social norms, of course, but much of our instincts and habits also provide information about our social world, not just our natural world. In collective rationalization, shared beliefs and desires are updated, but so are the overarching structures tying many beliefs and desires together. These collective cases involve what Cushman calls "deeper" forms of norm compliance: value caching and updating reward representations themselves.

This unique output of collective rationalization – the structures organizing many beliefs and desires into some coherent worldview or narrative – calls to mind the renewed academic interest in ideology, which Jost (2006, p. 652) defines as "any abstract or symbolic meaning system used to explain (or justify) social, economic, or political realities" (see also Graham & Yudkin, in press). Individual representational exchange can update specific beliefs and desires, while collective representational exchange updates the overarching structures organizing many beliefs and desires into shared meaning. For instance, work on system justification (Jost & Banaji 1994) has shown how ideological worldviews supporting existing social systems can predict collective action (Jost et al. 2017). Collective rationalization turns this on its head, as collective action becomes the input and the shared belief system becomes the output.

In collective rationalization, such outputs (belief systems, worldviews, narratives, and shared meaning) are especially likely to take on moral dimensions – moral belief systems, moral worldviews, moral narratives, and shared moral meaning. Group-level moral concerns like in-group loyalty and respect for authorities and traditions can be used to collectively justify the system (Haidt & Graham 2009) and bind groups into tightly knit moral communities (Graham & Haidt 2010). The group-binding function of these moral concerns can undoubtedly be advantageous for the group as a whole, and a shared sense of moral meaning can have palliative benefits for individuals in that group. But shared moral worldviews can also inspire violence within the group and against outgroups; for instance, acts of white supremacist hate crimes are often justified with language invoking moral imperatives of racial loyalty and purity (Graham & Haidt 2012).

Describing rationalization as a useful fiction, Cushman argues that "we can, should, and do actually adjust our beliefs and desires to match this fiction" (sect. 2.2, para. 6). Here "should" is meant to merely convey that doing so is adaptively advantageous (at least sometimes). But this is nevertheless a pretty rosy view of rationalization, and it is worth considering the moral "should" as well. Should Jason Bourne adopt the view that the man he just shot must have been a threat? Should a tribe or nation at war adopt the view that their enemies are evil and that God calls for their eradication? Collective rationalization highlights the dangers of representational exchange; rather than simple extractions of useful information from non-rational sources, these cases can involve motivated justifications or even denials of heinous past actions. For example, the ideology of manifest destiny in the United States has been described as a motivated rationalization of the atrocious collective actions of white immigrants, including slavery and Native American genocide (Winthrop 1852).

Cushman's account of rationalization centers on the true beliefs and adaptive desires this process can provide, and these of course often do co-occur. But collective rationalization shows how these two provisions can be in opposition: system justification and shared narratives about the moral exceptionalism of the in-group can be advantageous, in forming a loyal and orderly collective body, but these useful fictions come at the expense of the truth. The bias Cushman describes (in connection with individual theory of mind) to perceive all behavior as rational can motivate collective rationalization as well, as groups create shared beliefs and (objectively false) narratives to make sense of – and justify – their histories.

Collectives see what they've done and see it as justified. We drink our shared Kool-Aid together, and as history has shown, that can have disastrous consequences.

Cognitive dissonance processes serve an action-oriented adaptive function

Eddie Harmon-Jones and Cindy Harmon-Jones

School of Psychology, The University of New South Wales, Sydney, New South Wales, 2052 Australia.
eddiehj@gmail.com
cindyharmonjones@gmail.com
www.socialemotiveneuroscience.org

doi:10.1017/S0140525X19002176, e38

Abstract

The action-based model of cognitive dissonance proposes an adaptive function for rationalization that differs from the one offered by Cushman. The one proposed by Cushman is concerned more with the cold construction of cognitions, whereas the one proposed by the action-based model is a motivated protection of a strongly held cognition.

As Cushman notes, cognitive dissonance theory is "the best-known psychological theory of rationalization" (target article,

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

sect. 2.1, para. 3). Cushman follows up this bold statement by claiming that cognitive dissonance theory "does not offer an ultimate explanation – an answer to the question, "Why did it evolve?"" (sect. 2.1, para. 4) However, cognitive theory is not a unitary construct upon which everyone agrees. Since Festinger's (1957) original theory, a number of revisions and expansions have been proposed (for review, see Harmon-Jones 2019). One of these, the action-based model of cognitive dissonance, does offer an adaptive explanation for why dissonance evolved (Harmon-Jones & Harmon-Jones 2002; 2019; Harmon-Jones et al. 2009; 2015a).

The adaptive function proposed by the action-based model differs from the one offered in the target article, and this difference is likely due to the target article's incomplete understanding of cognitive dissonance processes and how they differ from self-perception-like processes. That is, most scientists working on these theories have agreed that these two theories apply to different psychological situations (e.g., Fazio et al. 1977; Harmon-Jones et al. 2019). Self-perception theory applies to situations in which organisms do not have strongly held perceptions, attitudes, or beliefs, and the self-perception process is concerned more with the *formation* or construction of perceptions, attitudes, or beliefs. In these situations, actions coldly cause rationalizations; the actions do not arouse negative affect, because the actions do not conflict with strongly held cognitions. In contrast, cognitive dissonance theory applies to situations in which organisms have strongly held perceptions, attitudes, or beliefs, and the dissonance process is one concerned with *protecting* one of those strongly held perceptions, attitudes, or beliefs. In these situations, actions motivate rationalizations because the actions aroused negative affect due to cognitive conflict (Fazio et al. 1977; Harmon-Jones 2000; Harmon-Jones et al. 1996).

The target article posits that the "function of rationalization, then, is to construct beliefs and desires that are consistent with the adaptive behaviors generated by non-rational processes, and then to adopt them" (sect. 3, para. 2). This psychological process is akin to a self-perception process, and we agree that this adaptive function of rationalization operates and applies to such self-perception-like situations. On the other hand, in dissonance-arousing situations, organisms are motivated to protect a strongly held "cognition" (e.g., perception, belief, emotion, behavior), and this cognition is the one to which the organism is most committed or regards as most true at the moment. In dissonance terms, this most strongly held, resistant-to-change cognition is known as the generative cognition, and cognitive changes will serve to protect or bolster this cognition. According to the action-based model, rationalization or protection of the generative cognition is posited to serve the adaptive function of enhancing effective action.

Why would rationalization be expected to enhance effective action? The action-based model posits that cognitions serve to guide action, that is, cognitions carry with them action tendencies. When cognitions conflict with one another, such that one cognition implies that another ought not to be true, their action tendencies are also likely to conflict. The organism will have a difficult time enacting a course of behavior when trying to follow the action tendencies evoked by conflicting cognitions. Thus, bringing cognitions more closely into alignment with the generative cognition makes the course of action clear, as the organism is not forever vacillating between alternative courses of action.

In line with the above, the action-based model posits that dissonance reduction is an approach-motivated process, and evidence supports this. For example, individuals who are higher in trait approach motivation engage in greater attitude change in the direction of a recent behavioral commitment (Harmon-Jones et al. 2011). Similarly, manipulating approach motivation, via cognitive or embodied means, influences the amount of cognitive dissonance reduction, measured as attitude change (Harmon-Jones & Harmon-Jones 2002; Harmon-Jones et al. 2015b). Furthermore, the degree of attitude change individuals engage in is positively related to their relative left frontal cortical activation, a pattern of neural responses that is associated with greater approach motivation (Harmon-Jones et al. 2008). These and similar results support the idea that dissonance reduction is a motivated process, not merely a matter of cold information-processing as in self-perception.

Additional support for the idea that dissonance was selected for by evolution is provided by evidence that non-human animals engage in "attitude change" in the direction of behavioral commitments. For example, capuchin monkeys show reduced liking for treats they had previously rejected (Egan et al. 2010). Similarly, pigeons show a preference for conditioned stimuli for which they had to exert more effort over those for which they had to exert little effort (Zentall & Singer 2007). As with humans, these changes in preference are biased in the direction of protecting an important cognition (a previous behavior or behavioral choice). The results in non-human animals suggest that dissonance processes are evolutionarily old and may be adaptive in a variety of species.

The target article emphasizes that rationalization may lead the organism to adopt (through non-rational processes) beliefs that are true. In contrast, the action-based model proposes that rationalization may be adaptive *whether the beliefs adopted are objectively true or not*. Simply reducing motivational conflict may allow individuals to behave more effectively, regardless of the objective truth or falsity of their beliefs. Truth can be a slippery construct, and it is often difficult to say with certainty whether a particular course of action is objectively better than another. We propose that it is often adaptive to fully commit to one's decisions and pursue them without hesitation, even if the alternate course of action may have been objectively slightly "better." Indeed, in cognitive dissonance research, rationalization often involves conflict between attitudes and motivations, rather than facts. Thus, we view the action-based model as a broader, more inclusive explanation for why dissonance processes evolved than that presented in the target article.

Acknowledgment. This work was funded in part by a grant from the Australian Research Council (DP180102504).

What kind of rationalization is system justification?

Kristin Laurin and William M. Jettinghoff

Department of Psychology, University of British Columbia, Vancouver, BC, V6T 1Z4 Canada.
klaurin@psych.ubc.ca
will.jettinghoff@psych.ubc.ca
https://magiclab.psych.ubc.ca/group/lab-director/
https://magiclab.psych.ubc.ca/person/will-jettinghoff/

doi:10.1017/S0140525X19002243, e39

Abstract

Cushman uses rationalization to refer to people's explanations for their own actions. In system justification theory, scholars use the same term to refer to people's efforts to cast their current status quo in an exaggeratedly positive light. We try to reconcile these two meanings, positing that system justification could result from people trying to explain their own *failure* to take action to combat inequality. We highlight two novel and contested predictions emerging from this interpretation.

In the target article, Cushman uses rationalization to mean something different from what we initially expected. He uses it to refer to people explaining their behavior: when one "takes an action that has already been performed and then concocts the beliefs or desires that would have made it rational" (target article, sect. 1.1, para. 1). In the system justification tradition, rationalization refers quite differently to people reinterpreting seemingly negative features of their sociopolitical systems, "adjusting their preferences to fit with their expectations about what is likely to occur" (Jost et al. 2004, p. 889; see also Laurin et al. 2013a; 2013b).

How can rationalization mean both explaining one's own behavior (per Cushman) and trying to feel better about the world one lives in (per system justification)? It could be a coincidence: Some words simply have multiple meanings. Here, we focus on two more fruitful possibilities.

Is system justification an instance of theory of mind plus social learning?

One possibility is that system justification comes from our tendency to infer (through theory of mind) – and then adopt (through social learning) – others' mental states. Cushman points out that this two-process sequence is exactly parallel to his meaning of rationalization: In each case, we observe (our own vs. another's) behavior, then infer and adopt the beliefs or desires that would have made it rational. System justification could be an instantiation of that two-process sequence: Justifying, for instance, gender pay gaps might be a by-product of trying to infer the rationale by which individual employers pay women less than men.

This account is plausible, but it does not explain why the extracted rationales so frequently tend to *justify* the system. System justification's main tenet would require that people be especially likely to infer in employers, and then adopt for themselves, beliefs and desires that *justify* the system or paint it in a positive light. For example, they should be especially drawn to extract the rationale that women must be less deserving of good pay. But if people were merely trying to identify beliefs and desires that would lead employers to *rationally* pay women less, there would be no reason to expect system-*justifying* rationales to predominate. Rationales that are neutral or even blame the system should be just as likely: For example, people could infer that women are less likely to retaliate than men in response to a mediocre offer. Thus, to account for the primacy of system-justifying beliefs, the theory of mind plus social learning requires additional assumptions that are not immediately obvious. We therefore explore an alternative that does not require such additional assumptions and also generates novel and contested predictions.

Is system justification the rationalization of one's own complacency?

This alternative is that people's tendency to rationalize undesirable features of society comes from their attempt to extract rationales that explain their own tacit support for society as a whole. In this account, people are not trying to explain the gender pay gap itself, but rather their own failure to fight it. This would account for the special appeal of system-justifying beliefs, particularly among the politically complacent (i.e., most humans; see Wike & Castillo 2018): By definition, system-justifying beliefs portray the status quo as more desirable, making action seem less necessary and complacency more rational.

Though consistent with the cognitive dissonance roots of system justification (Jost & Banaji 1994), this interpretation departs from most recent literature, which tends to imply that people are motivated to justify the system itself: They resolve dissonance between their preferences and a non-preferred reality with the ultimate goal of feeling good about the status quo (Bahamondes et al. 2019; Suppes et al. 2019). Instead, this new interpretation suggests people may be rationalizing their own tacit support for the system: They resolve dissonance between their inaction (and its implication that they prefer the status quo) and a non-preferred reality, striving to see themselves as rational actors. This leads to at least two novel predictions conflicting with assumptions in the current system justification literature.

When will people stop justifying the system?

Mainstream system justification literature predicts that people will only stop rationalizing the system when they believe it may change (Kay et al. 2002; Laurin 2018; Laurin et al. 2012). For example, if people expect gender inequality to shrink, then they should have no need to justify it, because their preferences and reality will soon be consonant.

By contrast, our Cushman-inspired reinterpretation of system justification predicts that people will stop rationalizing the status quo when they have attempted, even unsuccessfully, to change it, because the non-preferred reality is now consonant with their actions. For example, attending a feminist protest, even without any expectation that it would accomplish anything, should relieve any need to justify gender inequality, because one's actions are now in line with preferring gender parity.

Can complacency be explained without system justification?

The other novel prediction arises because our interpretation suggests a new alternative to system justification altogether: Rather than explaining their complacency by portraying action as unnecessary (i.e., by justifying the system), people might portray action as unfeasible. Accordingly, if we highlight the non-trivial costs of taking action against the system, that should release people from their tendency to justify it: If the costs of attending a gender equality protest (e.g., money lost in hourly wages, effort spent constructing a sloganized poster, etc.) explains why you have not attended a single one, then you no longer need to extract a belief that justifies gender inequality to explain your inaction. This prediction runs counter to what mainstream system justification research predicts: If anything, from that view, highlighting the costs of action should make the system seem even more static, which in turn should encourage people to justify it, rather than the opposite (e.g., Laurin et al. 2013a).

Conclusion

Applying Cushman's target article to system justification's version of rationalization leads to a new and parsimonious interpretation of system justification that generates two sets of contrasting predictions. Future research might test the following questions: (1)

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Will people stop rationalizing only when they believe attempts to change the status quo will succeed, or is merely participating in even an unsuccessful attempt sufficient? (2) Does highlighting the costs of attempting to change the system increase or decrease people's tendency to justify it?

Rationalization enables cooperation and cultural evolution

Neil Levy

Department of Philosophy, Macquarie University, Sydney, 2109 Australia;
Uehiro Centre for Practical Ethics, University of Oxford, Oxford OX1 1PT, United Kingdom.
neil.levy@mq.edu.au
https://mq.academia.edu/NLevy

doi:10.1017/S0140525X19002061, e40

Abstract

Cushman argues that the function of rationalization is to attribute mental representations to ourselves, thereby making these representations available for future planning. I argue that such attribution is often not necessary and sometimes maladaptive. I suggest a different explanation of rationalization: making representations available to other agents, to facilitate cooperation, transmission, and the ratchet effect that underlies cumulative cultural evolution.

Rationalization is usually thought of as a way in which we deceive ourselves or enable ourselves to perform wrongful or unwise actions while preserving a flattering self-image. There are, however, many instances in which we apparently engage in rationalization when neither moral nor prudential goods are at stake. Take cognitive dissonance (Cooper 2007). In the classic essay-writing paradigm, participants are either paid or requested (in a way that is difficult to refuse but leaves individuals feeling that they complied freely) to write counterattitudinal essays. Those participants who were requested to write the essays, but not those who were paid to do so, altered their reported beliefs in the direction of the claims defended in the essays. A natural explanation of these data is that we infer our own beliefs, in part, on the basis of our own behavior. While those who were paid could explain their behavior by reference to the financial inducement, those who were not paid inferred that they were committed to the view they defended.

Carruthers (2013) argues on the basis of evidence like this that belief attribution in the first-person case uses the same (observational) mechanisms as belief attribution in the third-person case. But if we are able to behave adaptively without attributing beliefs to ourselves, why do it at all?

Fiery Cushman suggests that such belief attribution is a useful fiction. When we rationalize our behavior, we extract information from it, taking ourselves to believe or desire whatever mental states best explain the behavior. This is a fiction when the behavior was caused by subpersonal mechanisms that lack such mental states. But it is a useful fiction, because these mechanisms encode responses that are adaptive. While our behavior is caused by mechanisms that are non-rational, beliefs and desires that *would* cause such behavior are themselves adaptive, and forming them allows us to rationalize – that is, make rational – future behavior. Rationalizing behavior allows us to transform their causes into explicit representations that then come to be available for planning.

The claim that explicitly represented information enables domain generality and thereby planning and flexibility of response is surely correct (Levy 2014). However, we should be wary of concluding that these gains explain the existence or yield the function of this kind of representational exchange. First, it is far from obvious that we need such beliefs and desires to engage in rational behavior. Cushman over-intellectualizes "reasoning," understanding it as inference over explicit representations. But there is a compelling case for understanding reasoning in a much less intellectualized way, as a flexible response to environmental and internal information (Levy 2019b). Cushman also over-intellectualizes "belief" and "desire," overlooking the fact that the great majority of these states are dispositional and may never be explicitly represented (see, e.g., Schwitzgebel 2002). Of course, we may use words however we want, but the substantive point worth emphasizing is that subpersonal mechanisms often drive flexible response in the absence of explicit representations. That being the case, it is unclear what we gain from extracting explicit information from them.

Second, the extraction of information may change future behavior for the worse. Consider cognitive dissonance again. When apocalyptic prophecies fail, cult members often become more committed to the cult, apparently to explain why they have devoted their resources to it (Dawson 1999; Festinger et al. 1956). We are quite easily manipulated into attributing beliefs to ourselves that may not serve our interests, as both cognitive dissonance and choice blindness experiments (Hall et al. 2012; Johansson et al. 2014) show. We confabulate differences between consumer products when the true explanation of our choices is mere position (Nisbett & Wilson 1977), thereby becoming willing to spend more on them in the future. In all these cases, we would do better to refrain from forming explicit representations and allowing subpersonal mechanisms to drive our behavior.

Perhaps the gains in planning and flexibility Cushman points to are among the functions of rationalization. At least equally important, however, is the capacity to make extracted representations available to *other* agents. We are deeply social animals, and our ecological success is very significantly due to our capacity to exchange and aggregate information (Henrich 2015; Richerson & Boyd 2008). To allow for such exchange, we often make ourselves transparent to one another (Funkhouser 2017). We evince a variety of signals that allow others to know what we're thinking. Language is, of course, our most powerful and flexible system of signals. An important part of the reason why we rationalize our behavior is not to make it rational – it already is – but to allow us to communicate its causes to others, thereby allowing them to make predictions about us and allowing them access to its contents for further epistemic work.

By making ourselves predictable, we enable more efficient cooperation, which is essential for the flourishing of social animals like us (Tomasello 2014). If each of us can predict how the others will behave, we can more efficiently play our part in joint actions, without interfering with one another or introducing redundancies. By communicating which aspects of our behavior are intentional, we indicate what should be copied and what

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

may safely be ignored, thereby facilitating the transmission of cultural knowledge (Levy & Alfano, 2019). Perhaps most importantly, making beliefs explicit allows them to be displayed to others. This both facilitates the "ratchet effect" that makes cultural evolution cumulative (Tennie et al. 2009), whereby previous innovations come to be a platform for further development, and also allows others to critically assess our representations to our epistemic benefit as well as theirs (Levy 2019a). The mechanisms of cultural evolution may work very much more powerfully when our beliefs are made explicit; given the importance of cumulative cultural evolution to our adaptive success, these gains are likely an important part of the explanation of the existence and function of rationalization.

Letting rationalizations out of the box

Philip Pärnamets[a,b], Petter Johansson[c] and Lars Hall[c]

[a]Faculty of Philosophy, New York University, New York, NY 10003; [b]Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, 171 77, Stockholm, Sweden and [c]Lund University Cognitive Science, Lund University, S-221 00, Lund, Sweden.
philip.parnamets@nyu.edu
https://philipparnamets.github.io
petter.johansson@lucs.lu.se
https://www.lucs.lu.se/choice-blindness-group/
lars.hall@lucs.lu.se
https://www.lucs.lu.se/choice-blindness-group/

doi:10.1017/S0140525X1900219X, e41

Abstract

We are very happy that someone has finally tried to make sense of rationalization. But we are worried about the representational structure assumed by Cushman, particularly the "boxology" belief-desire model depicting the rational planner, and it seems to us he fails to accommodate many of the interpersonal aspects of representational exchange.

In our work, we have studied rationalization using the choice blindness paradigm (Johansson et al. 2005). In a choice blindness experiment, participants make choices, the outcomes of which are surreptitiously manipulated to create mismatches to the participants' original selection. Participants often fail to notice this, and instead give detailed and coherent rationalizations for choices they never made. Moreover, research has shown that having participants accept false feedback about their responses and then rationalizing these responses can cause their attitudes to markedly shift in future choices (Johansson et al. 2014; Luo & Yu 2017; Strandberg et al. 2018), which seems like just the kind of adaptive construction of attitudes that Cushman posits. However, when trying to interpret our experiments in the light of representational exchange, we find a lot of blur in and between the central boxes of the theory.

It is not clear to us how the proposed theory handles the observed patterns of responses in experiments like ours. For example, some participants, while accepting the false feedback, later repeat their original decision, thus seemingly ignoring the self-generated arguments they just gave for the alternative option. Is it then these participants who best use the implicit information from their original decision mechanisms (prefer *X* not *Y*)? In other cases, some people in our studies, while accepting the false outcome, struggle to rationalize and sincerely say things like "I don't know/I'm not sure/I have no clue why *Y*." What should we make of these cases? If rationalization is rational, should we take these silent individuals, in a stunning reversal of previous canon, to be the irrational ones? In other words: When we attempt to take the theory seriously on its own terms, we seem to run into difficulties connecting it back to empirical results.

The blur between the modules is even better seen in Hall et al. (2012; 2013) and in Strandberg et al. (2018), where we have studied choices and rationalizations about moral and political choices. What stands out about these domains is the supposed involvement of the rational planning system. Clearly, people can harbor strong moral and political intuitions, which might be embellished and justified post hoc (as suggested by Haidt; 2001), but undeniably, our political and moral opinions are also the products of explicit rational argumentation, discourse, and thought (Cushman 2013; Rawls 1971). Thus, in these experiments we seemingly have the rational planning system rationalizing actions performed by itself.

But how could this work in proposed model, where the planner supposedly is the hub of all information flow? Could there really be a form of meta-rationalization about the rational plans we make? Or do we sometimes produce a form of "implicit" rationalizations, which are non-transparent to all the systems involved? In any case, Cushman needs to be more explicit about how his theory battles potential regresses and powerful homunculi piling on top of it, and about the localization of self and agency in all the potential layers of rationalization (Dennett 1991a). As far as we can see, the experimental examples described above do not fit any of the schemes for "hybrid" control described in the target article.

Broadening the perspective, it appears to us that Cushman's theory might underplay the complexity of the physical and social environment in which we act. Considering the old (but never aging) anti-representationalist slogan that "the world is its own best model" (Brooks 1991; Dreyfus 2007), it would appear that the results of our actions (telling us what we did) and the conditions motivating them (telling us why we did them) often will be evident directly in our immediate sensory environment, thus alleviating some of the need for internal information monitoring and exchange. Similarly, humans are embedded not only in a stable external world, but also in a social world. This world is imbued with a measure of stability by virtue of the various conventions (Lewis 1969/2008) and norms (Bicchieri 2005) that anchor our practices in a shared, interpersonal life-world (Wittgenstein 1953/2009; Von Uexküll 1934/2010). Our folk psychological practices are communal (Dennett 1991b; Wittgenstein 1953/2009) and indeed adapted to reason giving and coordination between agents – talk that invariably involves telling others (and ourselves) about what we are doing and why.

To us, this suggests an interpersonal component currently missing in the proposed theory, and one that differs from other proposals also emphasizing important aspects of the social grounding of rationalization (Mercier & Sperber 2011; Trivers 2000). Rationalization – and generally the casting of action into intentional language – might function by mediating representational exchange between agents. Agents construct meaning idiosyncratically as a function of their life histories (Freeman 1997), but need to share the hows and whys to form the communities and conventions that have allowed them to thrive. Perhaps some of the problems, sketched above, for the theory of representational exchange viewed as a flow of information only from certain decision systems to the rational planner, will dissolve if also taking into account the additional adaptive pressure of exchange between agents?

With this perspective, it becomes clear that we, and Cushman, need to consider the possibility of adaptive mismatches. If rationalization is an evolved adaptation, then it is an adaptation for a particular context and environment (EEA, or environment of evolutionary adaptedness; see Tooby & Cosmides 1992). But our modern environment of runaway information exchange and altered conditions for social interaction (group size, frequency, etc.) might differ considerably from the EEA. The power of the propositional attitudes of folk psychology (whether leveraged to understand ourselves or others) lies in compressing and abstracting information from the messy underlying systems (Dennett 1991c; 1996), but as a communal practice, it can also create too hard-edged opinions, attitude bloat, and distinctions where none are needed, particularly in a social context of conversational demands. Thus, a rationalization system that once faithfully abstracted useful information from our own and others' habitual, instinctual, as well as rational, actions now risks running in overdrive, with a real possibility that many day-to-day rationalizations are utter poppycock.

Rationalization is irrational and self-serving, but useful

Jake Quilty-Dunn

Faculty of Philosophy, University of Oxford, Oxford, United Kingdom, OX2 6GG;
Department of Philosophy, Washington University, St. Louis, MO 63105.
quiltydunn@gmail.com sites.google.com/site/jakequiltydunn/

doi:10.1017/S0140525X19002218, e42

Abstract

Rationalization through reduction of cognitive dissonance does not have the function of representational exchange. Instead, cognitive dissonance is part of the "psychological immune system" (Gilbert 2006; Mandelbaum 2019) and functions to protect the self-concept against evidence of incompetence, immorality, and instability. The irrational forms of attitude change that protect the self-concept in dissonance reduction are useful primarily for maintaining motivation.

Cushman usefully highlights the ubiquity and utility of rationalization processes. His account also reinforces an important and often-missed point: Rationalization provides a clear case of processing that is often unconscious but is nonetheless rationalistic, that is, involves a complex sense-making structure rather than mere association. The concept of representational exchange has many possible applications, offering to illuminate a wide array of phenomena beyond Cushman's key example of rationalization. This key example, however, does not in fact seem suited to illustrate the function of representational exchange.

Cognitive dissonance, in particular, does not have the function of increasing or repackaging knowledge about the world, ourselves, or our actions. Unconscious changes of attitudes to reduce dissonance are fundamentally *self-serving*. Aronson (1969; 1992) argued that dissonance reduction minimizes damage to the self-concept, which comprises the beliefs that we are good, that we are competent, and that we are stable. The negatively valenced feeling of dissonance arises when evidence contradicts these core beliefs, and rationalization occurs when we shift our attitudes to eliminate the contradiction in a way that preserves self-esteem.

The dissonance literature is full of such effects. Classic effects like effort justification (Aronson & Mills 1959) and the spreading of alternatives (Brehm 1956) arise out of a felt need to defang evidence of our own incompetence in decision making. Crucially, rationalistic dissonance reduction is modulated by self-esteem: For example, subjects who choose to shock a person who answers questions incorrectly assuage dissonance by reducing liking for that person, but not if their self-esteem has just been lowered (Glass 1964). More recently, the consumer research literature shows that meat eaters experience dissonance due to liking animals. If you're asked whether cows have emotions, having eaten beef jerky a moment before will lower your belief compared to subjects who ate cashews (Loughnan et al. 2010). The fact that I eat beef is not a good reason to think cows can't experience happiness, and the mechanism that underwrites this belief change does not seem to be a rationally good one.

Cushman pointedly asks what the "ultimate, adaptive" function of cognitive dissonance is. One answer that captures the irrational nature of dissonance reduction is: the preservation of stable motivation and avoidance of depression. Gilbert (2006) posits a complex *psychological immune system* to allow humans to stay happy and motivated in the face of adversity. Mandelbaum (2019) places dissonance, *qua* mechanism to avoid damage to the self-concept, at the center of this immune system. Why have a psychological immune system rather than pure rational updating? A perfectly rational updater saddled with human foibles and drives may not end up liking itself very much. The threat of depression and loss of motivation under such circumstances is stark. Cognitive dissonance is a drive state that pushes us to reorient our beliefs in ways that keep our self-esteem up, even when we are irrational, unstable, or immoral, and thereby keep ourselves motivated by avoiding the awful truth.

Viewing rationalization from the standpoint of the psychological immune system helps us make sense of evidence that is puzzling from the standpoint of Cushman's rationalism. For Cushman, rationalization helps us extract and repackage information about the mental causes of our actions. But dissonance effects can be quite distant from present actions and are filtered through our self-images in ways that pull us away from the truth. Merely being reminded of vegetarianism causes meat eaters to decrease belief in animal minds (Rothgerber 2014). Moreover, meat eaters are often driven to endorse a range of beliefs known as the "four Ns": meat-eating is necessary, normal, natural, and nice (Piazza et al. 2015). Meat-eating is tied to masculinity (Rozin et al. 2012), and men are correspondingly more likely to endorse the four Ns (Piazza et al. 2015). Watching a video detailing the animal suffering involved in meat production causes men not only to increase belief in the four Ns, but also to increase their commitment to eating meat (Dowsett et al. 2018). The fact that a practice is incorporated into a person's identity makes them more motivated to adjust their attitudes to let themselves off the hook when confronted with the immorality of that practice.

These cases of dissonance reduction seem geared more toward preserving our belief in our own moral goodness despite evidence to the contrary than toward the extraction of roughly accurate information about the world or ourselves. Without self-flattering rationalization, those of us who eat meat but dislike animal suffering would have to do the hard work of changing our diet or else come to the depressing realization that we are morally compromised. This problem applies quite generally. Apportioning beliefs about our actions and their

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

psychological origins to the evidence may cause us to downgrade beliefs in our own competence and morality, threatening emotional and motivational stability. And indeed, the "illusion of control" is more common among healthy patients than depressive patients, who are less likely to overestimate their own control (Alloy & Abramson 1979; Moore & Fresco 2012). These results tentatively suggest that rationalization can be undercut by depression, facilitating truth – and representational exchange – but endangering motivation (*modulo* the negative biases that also accompany depression; Beck 2008). Other unconscious processes of attitude change that aim at minimizing negative affect and preserving motivation through irrational means can also be found in the terror management literature (Pyszczynski et al. 2015). These various effects suggest a common immunodefensive function of avoiding negative affect and maintaining motivation in the face of one's flaws and inevitable demise, carried out through a variety of processes including dissonance-based rationalization.

Cushman's important target article helpfully refocuses attention on rationalization but neglects its self-serving nature. The unconscious rationalization processes underlying dissonance reduction function not to exchange information across systems, but to preserve motivation and avoid depression through protecting our image of ourselves by any irrational means necessary.

Rationalization of emotion is also rational

Peter Railton

Department of Philosophy, University of Michigan, Ann Arbor, MI 48109-1003.
prailton@umich.edu
https://lsa.umich.edu/philosophy/people/faculty/prailton.html

doi:10.1017/S0140525X19002292, e43

Abstract

Cushman seeks to explain rationalization in terms of fundamental mental processes, and he hypotheses a selected-for function: information exchange between "rational" and "non-rational" processes in the brain. While this is plausible, his account overlooks the importance – and information value – of rationalizing the emotions of ourselves and others. Incorporating such rationalization would help explain the effectiveness of rationalization and its connection with valuation, as well as raise a challenge to his way of bifurcating "rational" and "non-rational" processes.

Fiery Cushman brings a welcome synoptic vision to psychological processes and outcomes often considered in isolation from one another. By asking the evolutionary *why?* question, he brings attention to the need to understand how processes frequently associated with error could nonetheless have such a central and pervasive role in the human mind. This evolutionary perspective forces one to seek *deep* answers, since one must identify features of the mind that could plausibly be the result of natural selection, and so must look behind what plausibly are relatively recent appearances on the psychic landscape, such as rationalization's role in the enhancement of self-image or advertising of one's norm acceptance, which seem to require conceptual and linguistic capacities that probably appeared relatively late in evolutionary history. Moreover, the fact that rationalization as Cushman understands it – the post hoc construction of sense-making belief-desire explanations for one's own or others' behavior – figures in such a variety of mental phenomena suggests that it is the result of core features of mental architecture. Further, evolutionary explanation requires that one be able to point to benefits sufficiently large and reliable to favor such an architecture, despite its costs.

Therefore, at the heart of Cushman's account are processes that must have long evolutionary histories: reinforcement learning, causal modeling including inverse causal inference, and sense-seeking agential explanation. And the central benefit he identifies is one that could be real and important even when the content of a rationalization is false: information flow from non-rational to rational processes that can serve to improve reasoning and yield more successful outcomes.

This strikes me as very plausible, and I'd like to suggest generalizing the picture. First, a large part of everyday rationalization is of *emotion*, not just action: trying to make sense of our own feelings and moods, and those of others. On influential views of emotion and its evolution, affective states serve to synthesize multiple dimensions of information relevant to deciding how to act, predicting how others will act, and guiding one's behavior appropriately in context (Nesse & Ellsworth 2009). Yet because affective states bundle information together, making best use of them may require parsing, interpreting, and interrogating. If I feel uneasy in interacting with someone, *why?* Is it me or him or the interaction? What might this tell me about the situation, how I should react, how he's likely to react? Is the feeling appropriate, or am I over-reacting or prejudiced? And so on. Cushman's models of rational action (Fig. 1a in the target article) and rationalization (Fig. 1b in the target article) include beliefs and desires but not the agent's affective states as such, yet affect may be a primary source of the information needed to explain oneself and others, or to predict or evaluate what might happen next (cf. Thornton & Tamir 2017). At the same time, seeking to bring an emotional reaction into a rationalizing frame can contribute to *emotion regulation*, focusing or changing felt affect, and dampening dysfunctional effects (Wilson & Gilbert 2003). On the evolutionary side, it is not hard to imagine that capacities to read and regulate one's own emotions or moods by trying to make sense of them, or to seek to make sense of the emotions or moods of others as a way of informing one's interactions with them, would be, even if not perfectly reliable, capacities that facilitated everything from episodic cooperation to long-term relationships, and from parent-child relations and language learning to effectiveness in competition and warfare.

Second, taking affect into account might suggest a shift away from thinking of rationalization as exchanging information between "non-rational" and "rational" processes. Cushman uses "rational" both to designate "reasoning" processes and to pick out cognitive and choice processes that use evidence to form representations of probability and reward information and guide choice by maximizing expected utility. What, then, about mental processes that reliably perform the latter functions but are not forms of controlled reasoning? A large body of research suggests that the affective system contains such information-gathering processes and representational structures and guides decision making via comparisons of expected value (*modulo* the special role of risk; see Stauffer et al. 2014). Yet affect makes no appearance in Cushman's schematic models of our "multiple adaptive processes" (Fig. 2a) or of the operation of rationalization (Fig. 2b). No doubt the schemata in Figure 2 are meant to simplify these processes, but rationalization can be thought of as important for information sharing in the mind between controlled, deliberative "reasoning" processes and other kinds of processes, without thinking that this is primarily a matter of putting a rational mind in touch with non-rational processes. Emotions themselves are readily spoken of as more or less rational, and even habit may draw upon expected-value representations and weighings involving the "rational" affective system, rather than constituting a distinct "non-rational" system in itself (cf. Dayan & Berridge 2014; Smith & Graybiel 2016). This kind of mental integration seems even more likely for norm compliance. Greater mental integration can help with a problem posed by Hume that Cushman rightly considers at the end: if "rational" is restricted to "reasoning," then being "rational" won't always make us reasonable – it matters what our intrinsic values are, and affect is the currency of value. Neuroscientists have advocated abandoning the bifurcation of "cognitive" versus "affective" regions (Pessoa 2008). If rationalization includes fathoming and using the evaluative information that might be present in affect, as Hume (1738/1978) attempted in the *Treatise*, then we might find reasonableness in a less bifurcated picture of "rational" versus "non-rational" processes as well.

Rationalization and the status of folk psychology

Adina L. Roskies

Department of Philosophy, Dartmouth College, Hanover NH 03755.
adina.roskies@dartmouth.edu
https://philosophy.dartmouth.edu/people/adina-l-roskies

doi:10.1017/S0140525X19002139, e44

Abstract

Cushman's theory has implications for the philosophical debate about the nature of folk psychological states, for it entails realism about propositional attitudes. I point out a tension within his view and suggest a different view upon which rationalization emerges as a consequence of the adaptiveness of mentalizing. This alternative avoids the strong metaphysical implications of Cushman's theory.

Cushman has argued that rationalization is a form of representational exchange that has evolved because it is adaptive. Rationalization enables humans to change the format of information not initially encoded in propositional form into reasons, in the form of familiar constructs of folk psychology, such as beliefs and desires, that can positively affect future action. Cushman's view, if correct, has implications for classic and ongoing debates in the philosophy of mind about the metaphysical status of folk psychological states. There are three main views about the nature of folk psychology. Fodor (1980) has famously espoused a realist view of psychology, arguing that beliefs and desires (and hopes, fears, intentions, etc.) exist as propositionally structured entities in the head, and thus entail the existence of a language of thought. The representational and formal properties of these language-like states correspond to each other in such a way that the causally efficacious syntactic properties make possible reason-respecting behavior. Churchland's (1981) eliminativism is diametrically opposed to this Fodorian view: Churchland has argued that folk psychology is a mistaken and bankrupt theory, and that a better understanding of the brain will reveal that constructs such as beliefs and desires do not constitute natural kinds, that the representational structures that the brain uses are not language-like in any sense, but are rather far more articulated and richly multidimensional than anything that language can capture. Churchland used examples from connectionist research to illustrate his eliminativist arguments, and one might think that recent developments in the connectionist tradition, such as deep learning, may serve to vindicate his early arguments. Somewhere between these views lies the instrumentalism of Dennett (1987), whose position is that we can usefully characterize some systems as having beliefs and desires by taking a certain interpretive strategy, the intentional stance. To the degree that we can predict and explain a system's behavior by imputing folk psychological states to that system, so we are warranted in that imputation. For Dennett, even simple systems such as thermostats are legitimate targets of the intentional stance, even though no standard realist would support the view that thermostats have mental states. Dennett is a realist in that he claims there are real patterns in behavior that can be captured by the intentional strategy, but he remains agnostic as to whether intentional states map onto causally efficacious natural kind states that follow natural laws.

Although these classic debates in the metaphysics of mind seem far from the concerns of Cushman, these issues are deeply intertwined. Cushman argues that human rationalization is a process by which information from adaptive yet non-rational (in the sense of non-propositional) computational processes that drive behavior can be transformed in ways that make it accessible to rational processes. This species of informational exchange, moreover, encompasses a range of human cognitive functions: We call it rationalization when applied to our own behavior; we call it theory of mind when applied to the behavior of others.

If indeed rationalization – the process of metacognitively reframing observed behavior in terms of beliefs and desires – evolved because it was adaptive, then it must be the case that at least sometimes rationalization is causally efficacious, and the effects of rationalization have been overall beneficial. This seems to weigh heavily on the realist side of things. It does not entail that all behavior-causing states are folk-psychological states, but it does seem to entail that at least some of the states that cause behavior are precisely those emerging from the process of rationalization, which a fortiori are folk-psychological states. If so, it seems that a strong eliminativist position cannot be correct. It also seems that if eliminativism is correct – that is, if there are no classes of mental states that map cleanly to the propositional attitudes – then even if there is an important phenomenon of representational exchange, then whatever rationalizing amounts to, the propositional attitudes it yields cannot be the underlying causes of subsequent behavior. And if Cushman is correct, even the instrumentalist view seems overly quietist, for why favor a theory that refrains from speaking the truth?

While Cushman may be content to bite the bullet on metaphysical matters, there is nonetheless a tension in his argument. Cushman characterizes rationalization as (usually) a fiction, as it sees beliefs and desires as causes of our behavior where, by hypothesis, there are none. However, it is a useful fiction as it leads us to adjust our beliefs and desires to correspond to the fictional attributions, so that in subsequent action, we are driven by them. As a one-off account, this seems fine. But if rationalizing is as prevalent as Cushman indicates, and the propositional attitude-like outputs of the process affect behavior, then there is a constant looping: results of rationalization become causes of future actions, and rationalizations about these and subsequent actions do seem likely to identify the propositional attitudes that were causal, and thus they are not, by and large, fictions, but rather true descriptions.

If this tension makes Cushman's position seem unstable, one could find more plausible a slightly different story that nonetheless accepts the interesting parallelism between theory of mind and rationalizing. If social pressures were primary drivers of human evolution, as many have argued, one might think that mentalizing would be the primary adaptive phenomenon, and rationalizing may be the result of turning that adaptive tool upon oneself. On this version of the story, rationalizing could be a spandrel, and thus the story would not entail that rationalizing is rational or even beneficial (though it does not rule out the possibility). While the upshot for theorizing about representational exchange is unclear, on this alternative view, our ability to extract propositional attitude explanations from observing others' behavior does not entail realism about propositional attitudes. Indeed, it may be most congenial to a Dennettian view, for unlike Cushman's story, this story highlights the value of the intentional stance, rather than the underlying causal machinery of propositionally structured reasons.

Acknowledgments. Thanks to Brian Fiala and Jonathan Phillips for helpful discussions.

Rationalization: Why, when, and what for?

Rebecca Saxe[a] and Daniel Nettle[b]

[a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge MA 02139 and [b]Institute of Neuroscience, Newcastle University, Newcastle NE2 4HH, UK.
saxe@mit.edu saxelab.mit.edu
daniel.nettle@newcastle.ac.uk danielnettle.org.uk

doi:10.1017/S0140525X19002255, e45

Abstract

In this commentary, we ask when rationalization is most likely to occur and to not occur, and about where to expect, and how to measure, its benefits.

I can be puzzled by my own action. In this circumstance, I can't directly introspect my reasons, but I can infer plausible beliefs and desires from observing my actions. This is called rationalization. Cushman's central argument is that rationalization has a particular benefit: From rationalizing my actions, I learn true and useful information. The puzzling actions were generated by mechanisms that are not accessible to introspection, but are adaptive, serving my actual interests by increasing my capacity to survive and prosper. Thus through rationalization of my unplanned actions, I could come to make better future plans.

We have three questions about this idea.

When faced with another person's puzzling action, observers only sometimes rationalize it, assuming that the action was an efficient way to achieve the person's goals given her costs (Gershman et al. 2016). On other occasions, observers may explain puzzling behavior by writing it off as irrational. Indeed, observers too often infer that that people who chose unfamiliar actions don't just have different beliefs and desires, but are fundamentally irrational (Kennedy & Pronin 2008). Just as we can explain away others' actions as irrational, we can do the same for ourselves; I could interpret my own puzzling action as the result of a habit, or emotional aversion, or descriptive norm ("I always do that"; "it just feels bad"; "it's what we do"). So, our first question for Cushman is: Can we predict when observers do, and when they don't, rationalize (their own, or others') puzzling actions?

Perhaps we rationalize when the benefits of doing so are the greatest, but when is that? The target article provides surprisingly little empirical evidence of when and how the promised benefits of rationalization actually accrue. Cushman proposes that rationalization lets me learn useful new information, which I can incorporate into better subsequent plans. This idea ought to make testable predictions: Training or enhancing rationalization should cause improved planning; limiting or preventing rationalization should impair planning; and people who more frequently rationalize their past actions should therefore make better plans in future. Some phenomena do seem consistent with these predictions. For example, cognitive behavioral therapy, which may be a means of enhancing rationalization, is an effective strategy for reducing self-harm and improving problem solving (Hawton et al. 2016); and people who can give more distinct and differentiated descriptions of their emotions appear to cope better with unexpectedly negative or threatening events (Kashdan et al. 2010; Pond et al. 2012; Zaki et al. 2013). In general, we would like Cushman to translate his framework into more specific, falsifiable predictions.

On the other hand, even if some rationalization can generate some benefits for planning, more rationalization probably doesn't always generate more benefits (Trapnell & Campbell 1999). Too much rationalization is called rumination: Persistently chewing the cud of one's past actions is associated with depression and anxiety, not particularly good planning (Nolen-Hoeksema 2000). In sum, our second question for Cushman is: Under what circumstances do people who rationalize their actions actually make better plans? And how should we measure "better plans" to test this idea?

There may be a link here to the concept of "resource rationality" (Griffiths et al. 2015). The principle of resource rationality is to achieve the best cognitive outcome, counting the computational costs. A resource-rational creature would thus rationalize specifically when the benefits outweigh costs of more extensive computation; for example, where situations have complex and contingent payoff structures; where they involve close relationships of high value; around highly consequential decisions made under uncertainty; in the wake of conflict with significant others; or when following instincts, learned associations, or norms leads to a large error signal. It would be valuable to understand empirically if these are the cases where most rationalization does indeed occur. The idea of resource rationality also links quite naturally to the possibility of pathological over-rationalization:

This would occur when rationalization resources are deployed beyond the point where their marginal benefit for future situations could exceed their cognitive cost.

Finally, we feel the target article may underestimate the importance of the interpersonal functions of rationalization, as well as how tight the linkage is between inter- and intra-personal ones. Cushman considers one way in which rationalization operates in interpersonal interaction: when I offer an exculpatory but false explanation of harmful actions. In this context, since I know my explanation is false, the only reason I might actually acquire the rationalized beliefs and desires is to achieve more effective deception of others (Trivers 2000). We suspect that this example is too narrow, missing much of the proximal motive and the ultimate benefit of rationalization.

Like Cushman, we expect that human minds contain multiple mechanisms vying for control of behavior. In our view, such competition occurs not just between separate modules with different functions. There can also be conflict within any one system, for example, between competing desires, fears, norms and expectations. Inner conflicts arise because people are uncertain about the present, about the future, about the value of key rewards, and so on.

Our third question is therefore: do the benefits of rationalizing one's own actions actually arise mostly from simplifying the sources of our actions that we share with other people? When a person's actions actually emerge from inner conflict and competition, the resulting sequence of behaviors could be quite confusing to third-party observers trying to infer a unitary set of beliefs and desires. Rendering our own actions seemingly rational and comprehensible could have many interpersonal benefits, including signaling our value as a future partner or persuading others to subscribe to joint plans (Mercier & Sperber 2011).

In short, the target article stimulates further questions: about when rationalization is most likely to occur and to not occur, and about where to expect, and how to measure, its benefits.

Antecedent rationalization: Rationalization prior to action

Eric Thomas Sievers

Department of Philosophy, Florida State University, Tallahassee, FL 32306-1500.
esievers@fsu.edu www.ericsievers.xyz

doi:10.1017/S0140525X19002188, e46

Abstract

Often times we find ourselves wrestling with the urge to commit a non-rational action. When this happens, we are quite good at adopting quasi-beliefs that, if true, would make the action rational. In other words, rationalization often occurs antecedent to a behavioral choice. A complete account of the evolutionary history of rationalization must include antecedent rationalization.

Imagine you are ordering a salad at your favorite deli. This dietary choice reflects your desire to lose weight and an accompanying belief that a reduction in daily caloric intake is necessary to accomplish your goal. The young man at the counter innocently inquires if you would like to add bacon to your salad. You begin trying out various beliefs that, if true, might allow you to eat the bacon in good faith. In other words, you begin rationalizing an action that you know to be contrary to your overarching desires and beliefs before beginning that action. You order the bacon and commit to an unfounded justificatory belief-like state against your better epistemic sense.

While the above may not be an exciting enough scene to warrant inclusion in the next Jason Bourne film, it demonstrates an aspect of rationalization that is neglected in Fiery Cushman's account – rationalization may occur antecedent to a behavioral choice (sect. 1). It is not only a device which "concocts the beliefs or desires that would have made [a past action] rational" (sect. 1.1, para. 1). Rather, rationalization often crafts such beliefs, or rather some temporary justificatory analog to belief, in order to allow ourselves to perform an action driven by a non-rational process. Ignoring antecedent rationalization obscures our investigation into the nature of rationalization generally. As such, my commentary will argue for its inclusion in an account of the overall evolutionary history and function of rationalization. While I do not reject Cushman's conclusion that rationalization is an example of a representational exchange mechanism, my claim here is that an account of rationalization that does not include antecedent rationalization is incomplete.

Cushman points out that rationalization is a useful fiction because it can improve subsequent reasoning by crafting post hoc reasons for our non-rational actions. In the case of antecedent rationalization, a different sort of useful fiction emerges from an urge to act in a non-rational way. Picture our Pleistocene ancestor battling a desire to behave in some non-rational fashion. Perhaps she is considering concealing a portion of some collective resource in violation of a group norm. She wants to stash away a bit of food to eat later when she is alone. Assume for the moment that this is a non-rational decision because being caught would lead to a loss of trust between her and other group members and that this will inhibit her fitness in the future. Assume also that getting caught is likely and the small payoff that the snack will provide is not worth the risk. She begins to adopt various quasi-beliefs that would allow her to carry out her plan. First, she underestimates the likelihood of being caught. Then she overestimates the likelihood of other group members engaging in similar behavior. She eventually concocts a useful fiction that allows her to steal away her extra portion of food against her better judgment and in direct defiance of her overarching beliefs and desires concerning her fitness optimization. She rationalizes a non-rational behavior prior to engaging in it. We can imagine a similar process occurring prior to more contemporary examples of non-rational behavior: the adulterous spouse, the tax evader, even the murderer. Of course, as in the case of cheating on your diet, antecedent rationalization does not only precede norm violations. The resulting behavior does not even need to be negative. It may be irrational to drop out of a university program to pursue a career as an artist. One can imagine some amount of belief manipulation is necessary to commit to such a risk. The point is, in all the cases above, someone mixes their own Kool-Aid before the party begins, just so they can indulge in sipping it (sect. 1.1, para. 4).

Nothing entailed by antecedent rationalization precludes the correctness of Cushman's model of rationalization as a representational exchange mechanism. It does, however, demand an expansion of that model. Antecedent rationalization is consistent with Cushman's conception of seeking cognitive consonance as well as the sort of reflective equilibrium he describes in the context of moral reasoning (sects. 1.1 and 4.5). It is consistent with the notion of non-rational action being guided by instinct, habit, and norms (sect. 4.1). It squares with the functions of impression management and responsibility avoidance that Cushman attributes to rationalization (sect. 2.3). It may function in part, for example, to provide a pre-social explanation to have at the ready in case a nefarious actor is found out. Alternatively, perhaps it emerged in part because gambling pays off just often enough to allow for the preservation of a mechanism that rationalizes risky behavior. In any case, it is consistent with Cushman's representational exchange model and possesses sufficient overlap with the functions of rationalization to which he draws attention. In short, antecedent rationalization is a form of rationalization. As such, Cushman must modify his model in a way that accommodates it.

It may be argued that a justificatory belief-like state is not a belief. To state this is to ignore their relevant functional similarity. They both justify non-rational behavior. In both instances, we piece together what would need to be believed in order to make a particular non-rational action a rational one. We then behave in a manner consistent with those beliefs being true. In both cases, we stir up a useful fiction. Alternatively, it could be argued that what I'm describing as antecedent rationalization is just a form of rational planning. The distinction, however, is that in the case of antecedent rationalization, the actor knows that the behavior is contrary to his overarching beliefs and desires yet guides his decision by fictionalized quasi-beliefs.

Antecedent rationalization is not a counter-example of Cushman's model. It fits neatly within the concept of rationalization as a representational exchange mechanism. It is, however, a ubiquitous form of rationalization that is absent from Cushman's treatment. As such, Cushman's model must be expanded to include and account for antecedent rationalization.

Ex ante coherence shifts

Dan Simon[a] and Keith J. Holyoak[b]

[a]Gould School of Law and Department of Psychology, University of Southern California, Los Angeles, CA 90089-0071 and [b]Department of Psychology, University of California, Los Angeles, CA 90095-1563.
dsimon@law.usc.edu
https://weblaw.usc.edu/faculty/?id=307
holyoak@lifesci.ucla.edu
https://www.psych.ucla.edu/faculty/page/kholyoak

doi:10.1017/S0140525X19002103, e47

Abstract

Cushman characterizes rationalization as the inverse of rational reasoning, but this distinction is psychologically questionable. Coherence-based reasoning highlights a subtler form of bidirectionality: By distorting task attributes to make one course of action appear superior to its rivals, a patina of rationality is bestowed on the choice. This mechanism drives choice and action, rather than just following in their wake.

Cushman argues that after engaging in behavior triggered at least in part by processes other than thinking (e.g., instincts and habits), people infer reasons that rationalize their behavior. These rationalizations, in turn, serve to adjust people's beliefs and desires to help guide future decisions and behaviors. This defense of the rationality of rationalization is insightful, but in our view, its scope and contribution are limited in two important ways.

First, the analysis posits that rational behavior is driven straightforwardly from "a mechanism that chooses actions by maximizing the satisfaction of your desires, given your beliefs" (target article sect. 3.1, para. 1). Thus, the model glosses over situations in which the putatively rational components – namely, desires and beliefs – do not align in perfect harmony. Yet such non-alignment is typical of consequential decisional dilemmas that people face in everyday life. Consider the numerous conflicting, uncertain, and incommensurable factors involved in choosing which college to attend, whether to accept an apology for an insult from a colleague, or even just deciding what to order from the food truck. A growing body of research shows that people reach decisions in such dilemmas neither by rational forward inference nor by backward-oriented rationalization. Rather, a subtler form of reasoning distorts the task attributes to make one course of action emerge as dominant over rival choices.

In models of coherence-based reasoning (Holyoak & Simon 1999; Read & Simon 2012; Simon 2004; Simon & Holyoak 2002; Spellman et al. 1993), complex decision situations are represented by networks in which the relevant variables are interconnected via excitatory and inhibitory links. Constraint satisfaction mechanisms settle the network into a stable state of coherence, in which mutually supportive connections (i.e., those that "go together") activate one another and collectively inhibit their rivals (cf. McClelland & Rumelhart 1981). Thus, variables in the network come to cohere with the emerging decision: Those that support the winning conclusion are strongly endorsed, whereas those that support the rejected interpretation are suppressed or rejected (Holyoak & Thagard 1989; Read et al. 1997; Thagard 1989). This spreading-apart transforms difficult choice dilemmas into obviously correct decisions and thus serves the adaptive goal of affording confident choice in the face of decisional conflict (Glöckner et al. 2014; Janis & Mann 1977; Read & Simon 2012; Simon & Holyoak 2002). This form of reasoning does not typically lead to irrational decisions (as in preference reversal); more often, it exaggerates small initial preference differentials into large disparities, thus bestowing a patina of rationality on the eventual choice.

Coherence-based reasoning has been demonstrated across a wide range of decisional domains, including social judgment (Simon et al. 2015), legal reasoning (Holyoak & Simon 1999; Simon et al. 2001), moral reasoning (Holyoak & Powell 2016), evidence evaluation (Carpenter et al. 2016; Engel & Glöckner 2013; Simon et al. 2004b), probabilistic judgments (Glöckner et al. 2010), choice between jobs (Simon et al. 2004a; Simon & Spiller 2016), consumer choice (Simon & Spiller 2016), and attitudes toward war (Spellman et al. 1993). Similar predecisional distortion of task attributes has been observed in other research traditions, spanning consumer decision making (Chaxel et al. 2013; Russo et al. 1996; 2008), evidence evaluation (Carlson & Russo 2001), medical decision making (Kostopoulou et al. 2012), and risky decision making (DeKay et al. 2009).

Second, Cushman's analysis posits that rationalization always follows action (as illustrated in Fig. 2 in the target article). This temporal ordering echoes Festinger's (1957; 1964) insistence that cognitive dissonance arises only after the person has engaged

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

in counter-attitudinal behavior or has committed to an imperfect choice (see also Brehm 1956). This viewpoint was challenged almost immediately. Bruner (1957) criticized the ex post facto nature of dissonance theory, claiming that "the most interesting aspects of cognition are those that precede the making of decisions rather than those that follow [it]," adding that the theory amounted to a "rather autistic tradition" (p. 152). Abelson (1983) lamented that cognitive dissonance theory had been reduced to explaining how people "recover from experimentally engineered major embarrassments" (p. 43). Abelson insisted that the behavior of people who have been forced to make "damned fools of themselves" is a distraction from the broader potential applications of structural dynamics (see also Berkowitz & Devine 1989). As Heider (1979) argued, the tendency toward consistency should not be interpreted as merely repairing disturbances of balance; rather, it implies reaching out to bring the various pieces of the cognitive field into consonance. Indeed, research findings indicate that coherence-driven distortions largely occur *prior to* the time at which the decision is made or an action taken (DeKay 2015; Holyoak & Simon 1999; Russo et al. 1998; Simon et al. 2001; for reviews, see Brownstein 2003; Simon & Holyoak 2002).

A nice illustration of subtle bidirectional reasoning is provided by the reflective equilibrium model (Daniels 2003). As discussed in the target article, when making a considered judgment, the reasoner seeks to harmonize principled rules or reasons with their intuitive judgment about the particular case. When the two do not naturally coincide, the thinker modifies them both to bring them into alignment at a state of equilibrium. Reflective equilibrium is typically considered to be a conscious process, whereas coherence shifts appear to occur mostly beneath conscious awareness (Holyoak & Simon 1999); nonetheless, all these processes involve bidirectional reasoning that generates an optimal choice and action given the constraints of the situation.

In sum, by focusing exclusively on the construction of ex post facto rationalizations, Cushman's analysis fails to capture the ex ante impact of bidirectional reasoning on human decision making. We contend that the scope of coherence-based reasoning is far broader, and addresses more important cognitive challenges, than does simple rationalization to justify decisions already reached and actions already taken.

Evidence for the rationalisation phenomenon is exaggerated

Tom Stafford

Department of Psychology, University of Sheffield, S1 2LT Sheffield,
United Kingdom.
t.stafford@sheffield.ac.uk
http://tomstafford.staff.shef.ac.uk/

doi:10.1017/S0140525X19002085, e48

Abstract

The evidence for rationalisation, which motivates the target article, is exaggerated. Experimental evidence shows that rationalisation effects are small rather than gross and, I argue, largely silent on the pervasiveness and persistence of the phenomenon. At least some examples taken to show rationalisation also have an interpretation compatible with deliberate, knowing reason-responsiveness on the part of participants.

The evidence for rationalisation, which motivates the target article, is exaggerated. There are two sources for this. First, it is an outcome of structural features of the experimental psychology tradition, which isolates effects using experimental control, and then uses null-hypothesis significance testing to establish their reality (i.e., their non-zero size), neglecting to gauge their importance (Stafford 2014). The second source of this exaggeration is the rhetoric of psychologists, who make hay out of emphasising the supposedly irrational aspects of our behaviour, de-emphasising reason-responsiveness (Stafford 2015).

With respect to the target article, we can see this in the motivating section (1.1. "Rationalization"). The concept of rationalisation is presented as pervasive ("people rationalize all the time"), solidly evidenced ("exhaustively documented"), and leading to "gross errors," which are "stubbornly irrational."

A study recruited as a key illustration of rationalisation is Sharot et al. (2010), but when we look at this example, we see that the largest effect reported in this paper was an average within-subject change of ∼0.07 on a 6-point scale (experiment 1, see Fig. 1, $t(20) = 2.4$, $p < 0.03$), so the rationalisation manipulation produced a mean shift of ∼1% in people's judgements. Hardly gross or stubborn.

Other examples cited by the target article are similar – showing small movements in people's ratings of belief, rather than flips from one belief state to another. This contrasts with the rhetorical portrayal of rationalisation. Participants in Brehm's (1956) study showed an average change of ∼0.9 on an 8-point scale (i.e., an ∼11% between-groups shift, $p < 0.01$, with $n = {\sim}30$ in each group). The experiment design, analysis, and presentation of results presented by Vinckier et al. (2019) does not make a simple estimation of effect size for rationalisation obvious, but according to their Fig. 3, it looks like the within-subject effect size of choice (i.e., of rationalisation) is ∼0.15 of a standardised (z) score. A "small effect," as classically determined (Cohen 1992).

Some might view this as unproblematic. A reasonable view is that rationalisation is indeed common and commonly produces "gross errors," but the methods of experimental psychology mean we can only hope to consistently capture rationalisation in the proxy form of meagre shifts on rating scales. But it is also a reasonable view, I believe, that the extant evidence for rationalisation does not support the grand claims for the power of the phenomenon. It is not enough that an idea be intuitively plausible.

The target article also invokes classic and widely known studies in psychology as evidence of rationalisation. It is not possible to review all of them in this commentary, but it is instructive to look at couple of salient examples. Cognitive dissonance is the first citation of the target article (Festinger 1962). The foundational demonstration of cognitive dissonance is Festinger and Carlsmith (1959). Taking the largest effect reported in this study, participants' ratings of "enjoyability of task," this showed a mean shift of 1.40 on an 11-point scale (i.e., ∼13% change between groups, $p < 0.03$, $t(38) = 2.22$, with $n = 20$ in each group).

Another foundational contribution is Nisbett and Wilson (1977), for which (surely) the most discussed study is Wilson and Nisbett (1978, experiment 2), in which shoppers were asked to choose between four identical pairs of stockings. Famously, shoppers preferred the stockings on the right ($p < 0.025$, $n = 52$), but gave reasons other than position for their preference. Here my issue is not the size of the effect, but of its interpretation. While the explanations offered could be due to rationalisation, failure to report reasons is not the same as their inaccessibility. Stafford (2014) argues that, where participants are ignorant of the conditions by which experimenters analyse their behaviour, failure to report those conditions is wholly compatible with rational choice.

Note my argument is not about the reproducibility, or not, of the evidence base, but rather of its interpretation.

It is telling that other motivating examples evoked by Cushman are fictional and/or psychologically exceptional (amnesia in the *Bourne Identity*, split-brain patients). While these are highly suggestive of the human potential of rationalisation they cannot be taken as evidence that rationalisation is pervasive in ordinary cognition.

I have argued that errors due to rationalisation are often small, rather than gross. The interpretive slippage between experimental effects at single time points and unusual edge cases (like split-brain patients) means that the evidence for the persistence ("stubbornness") of rationalisation is simply unclear. Further, while the target article notes that reasons are a factor in driving action, alongside rationalisation, it is possible that some evidence presented as demonstrating rationalisation could actually be showing rational behaviour (so, for example, in Wilson & Nisbett 1978, it could be viewed as rational to give and defend an implausible answer if a psychologist asks you an impossible question such as which pair of identical stockings you prefer).

Where, then, does this leave the Cushman "representational exchange" account? I am not claiming that rationalisation doesn't exist, only that the evidence psychology has produced in support of it is far less strong than is commonly supposed. There is still – potentially – something for Cushman's account to explain. My criticism highlights an opportunity: One test of the value of Cushman's account is if it can provide experimentalists with the leverage to produce better evidence of rationalisation; that is, the value of the account could be in allowing us to predict when and how rationalisation will be at its strongest and most divergent from simple rationality in which actions are motivated by consciously accessible reasons.

Perhaps, guided by better theory, experimentalists will be able to generate manipulations that show rationalisation in the lab consonant with our intuitions of its importance in our everyday lives.

Rationalization may improve predictability rather than accuracy

P. Kyle Stanford[a], Ashley J. Thomas[b] and Barbara W. Sarnecka[c]

[a]Department of Logic and Philosophy of Science, University of California, Irvine, CA 92697-5100; [b]Department of Psychology, Harvard University, Cambridge, MA 02138 and [c]Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100
stanford@uci.edu
https://faculty.sites.uci.edu/pkylestanford/
athomas@g.harvard.edu
https://ashleyjothomas.wixsite.com/mysite
sarnecka@uci.edu
https://sites.google.com/uci.edu/sarneckalab

doi:10.1017/S0140525X19002279, e49

Abstract

We present a theoretical and an empirical challenge to Cushman's claim that rationalization is adaptive because it allows humans to extract more accurate beliefs from our non-rational motivations for behavior. Rationalization sometimes generates more adaptive decisions by making our beliefs about the world less accurate. We suggest that the most important adaptive advantage of rationalization is instead that it increases our predictability (and therefore attractiveness) as potential partners in cooperative social interactions.

Cushman makes a compelling case that rationalization is a form of what he calls "representation exchange" and that such representation exchange is itself a crucial feature of action-guiding cognition. We agree with Cushman that these processes of representation exchange, including rationalization, are broadly adaptive. We are less convinced, however, by his claim that rationalization is adaptively advantageous because it allows humans to extract valuable information ("true beliefs and useful desires") from the highly evolved inclinations of our non-rational action-guiding systems. We see two problems with this proposal, one theoretical and one empirical.

We begin with the theoretical problem. Cushman recognizes that our actions are generated by a complex combination of rational (beliefs, desires) and non-rational (instincts, habits, norm compliance) impulses, and he suggests that rationalization "constructs new beliefs and desires where none had existed, to extract information from the *non-rational* processes that influence our behavior" (target article sect. 1.1, para. 7). Thus, if I must decide which route to take as I walk my dog, and the complex product of the non-rational and rational influences on my behavior leads me to walk along the road rather than along the river, I might rationalize that decision by deciding that I am afraid of the river route. On Cushman's account, I will then come to accept and adopt that rationalization (my fear of the river route or a belief about its danger) as a further, conscious motivation for avoiding the river. But creating this new, additional motivation will not change the impact of those same non-rational action-guiding impulses on my behavior. Thus, if I face the same (or a similar) decision tomorrow, the very same non-rational action-guiding impulses will remain in place and continue to influence my behavior, ensuring that my degree of aversion to the river route is not explained by the magnitude of my new consciously accessible fear (or my beliefs about its danger) and triggering a new round of rationalization that strengthens or intensifies that fear in response. (Indeed, if these non-rational impulses fully preserve their influence, the degree of mismatch between my actions and my consciously accessible motivations should remain just as large as it was initially.) It seems that this process of repeated rationalization will continue as long as the relevant behavior is generated by a combination of non-rational and rational impulses, ultimately ensuring that the accumulated wisdom of my non-rational action-guiding impulses is substantially *overrepresented* in our actual (complex, multisystem) decisions about what to do (e.g., leaving me absolutely terrified of taking the river route). This would not be cause for concern if we thought the guidance of our non-rational action-guiding systems were always correct, but in that case, there would be no evolutionary advantage in making such decisions accessible to rational influence in the first place.

The empirical problem is that even if rationalization invariably produces more fitness-maximizing choices and decisions, it sometimes does so by giving us less accurate beliefs about the world. One example is Liu and Ditto's (2012) work on moral coherence: Briefly, subjects who are induced by argument to shift their views concerning the *moral defensibility* of the death penalty will also shift their views about the extent to which the death penalty is *practically effective* in deterring crime. And we ourselves have shown that subjects will judge a child left alone in precisely the same circumstances to be in significantly more danger if the parent leaves for a morally unacceptable reason (e.g., an adulterous affair) than a morally neutral reason such as going to work (Thomas et al. 2016). In these studies, it seems that representation exchange is inducing subjects to modify their factual beliefs in ways that do not increase their accuracy but instead make them better cohere with the subjects' own moral judgments. This may well ensure more adaptive responses (ones that better reflect the community's normative views, for example), but if so, this is achieved by making the subject's beliefs *less* accurate or responsive to relevant evidence.

The challenges we have presented also suggest an alternative to Cushman's hypothesis concerning the adaptive benefits of rationalization itself. Even if it does so at some cost to the accuracy of our beliefs, the one thing representation exchange undeniably increases is the coherence of our motivations for action and (therefore) the predictability of our behavior to others. There is reason to think that behaving predictably has adaptive value in its own right, as emphasized by those evolutionary theorists who appeal to mechanisms of partner choice, reputation management, and the virtues of predictable social partners more generally in seeking to understand the evolution of human ultrasociality (e.g., Baumard et al. 2013; Stanford 2018; Tomasello, in press). Thus, while we agree with Cushman that representation exchange is broadly adaptive and will sometimes increase the accuracy of our beliefs, we doubt that this is the only or even the most important way in which rationalization in particular increases our fitness. We suggest that rationalization in humans is adaptive in large part because it renders us more predictable (and therefore more attractive) partners for one another in the sorts of hypercooperative social structures on which human societies depend.

Quantifying the prevalence and adaptiveness of behavioral rationalizations

Warren Tierney[a] and Eric Luis Uhlmann[b]

[a]Kemmy Business School, University of Limerick, Castletroy, Limerick V94 T9PX, Ireland and [b]Organisational Behaviour Area, INSEAD, 138676 Singapore.
warrentierney@hotmail.com
eric.luis.uhlmann@gmail.com http://socialjudgments.com/

doi:10.1017/S0140525X1900205X, e50

Abstract

Critical aspects of the "rationality of rationalizations" thesis are open empirical questions. These include the frequency with which past behavior determines attitudes (as opposed to attitudes causing future behaviors), the extent to which post hoc justifications take on a life of their own and shape future actions, and whether rationalizers experience benefits in well-being, social influence, performance, or other desirable outcomes.

Cushman posits that rationalization of past behaviors extracts information from non-rational psychological processes such as habits and instincts, rendering it available for later reasoning. In doing so, post-hoc rationalizations improve the reasoning processes that follow.

We suggest that key aspects of the rationality of rationalizations thesis are open empirical questions, among these the prevalence of behavioral rationalizations, the extent to which rationalizations are carried over to future judgments, and whether rationalizations lead to desirable outcomes for the person engaging in them. Such empirical questions can be addressed through studies capturing dynamic interactions between self-reported attitudes and behaviors over time, as well as the correlates and downstream consequences of behavioral rationalizations.

How prevalent a phenomenon are behavioral rationalizations, in other words cases in which past behaviors determine future explicit attitudes?

The available longitudinal evidence suggests that Time 1 explicit attitudes predict Time 2 behaviors far better than past behaviors predict future self-reported attitudes, calling into question the prevalence of post hoc rationalizations for past actions (Bentler & Speckart 1981; Fredricks & Dossett 1983; Kahle & Berman 1979). Popular perspectives on attitude-behavioral relations may be "surprise-hacked" (Felin et al. 2019), overemphasizing instances in which behaviors cause explicit preferences (Bem 1972; Festinger 1962), and automatic and unintentional processes determine human behavior outside of conscious awareness (Caruso et al. 2017; Forscher et al. 2019; Lodder et al. 2019; McCarthy et al. 2018; Oswald et al. 2015). Although further longitudinal and meta-analytic investigations are needed, the "boring" narrative that conscious preferences and intentions typically direct future actions may capture a far greater share of the variance (Armitage & Conner 2001; Ajzen 1985; Fishbein & Ajzen 1975; Randall & Wolff 1994; Sheppard et al. 1988; Webb & Sheeran 2006), relegating the rationalizations-are-rational thesis to address only a small portion of the attitude-behavior relationship.

Once formed, are rationalizations carried over to future judgments?

In other words, do explicit preferences formed to justify past acts play a causal role in directing future actions, or are such conscious rationalizations brief coping mechanisms, or a mere residue of behaviors determined by implicit processes (Gazzaniga 1985)? One relevant experiment on moral judgments manipulated victim race, finding that whether the individuals sacrificed are White Americans or Black Americans impacts if consequentialist versus deontological values are endorsed as general principles (Uhlmann et al. 2009). Further, once formed, such motivated moral principles impact downstream judgments. For example, if deontological morality is endorsed in a motivated fashion because the victims are Black Americans in the first moral dilemma, the same principle is then applied to a second moral dilemma in which victims are White Americans. Although further studies testing for such carryover effects are needed, this provides initial evidence that rationalizations can play a causal role in future judgments, a key aspect of Cushman's thesis.

At the same time, the Uhlmann et al. (2009) results and related findings on intergroup attitudes (e.g., Brescoll et al. 2013; Hodson et al. 2002; Norton et al. 2004; Tannenbaum et al. 2013) seriously question whether rationalizations improve subsequent reasoning. For example, individuals who exhibit negative automatic associations with overweight people on indirect measures are also more likely to explicitly favor increased insurance premiums for overweight employees. Yet they justify such punitive policy preferences in terms of cost effectiveness, rather than personal beliefs about body weight (Tannenbaum et al. 2013). Given that target ethnicity and obesity are not defensible inputs into moral judgments in the first place, how does rationalizing group-based biases and then carrying forward such justifications improve subsequent reasoning in any way? Even assuming for a moment that implicit preferences are somehow "truer" or more authentic than explicit preferences (a highly debatable characterization), the rationalization process has obscured, rather than revealed, this deeper attitude. Applying the criterion of subjective rationality (Pizarro & Uhlmann 2005), it seems doubtful that decision makers themselves would, if made aware of it, welcome the influence of implicit overweight bias on their recommended company insurance policies. More likely, we think, they would seek to correct for and remove such unwanted prejudices (Fazio 1990) and perceive them as in conflict with their ideal self (Monteith et al. 1993). This leads us to the broader issue of whether post hoc justifications are "good" for the rationalizer in some measurable way.

Do rationalizations lead to positive objective or subjective outcomes for the agent?

If the "ultimate purpose of reasoning" is "fitness maximization," and rationalizations improve reasoning (target article, sect. 2.1, para. 9), then individuals who engage in rationalizations should score higher on measures of adjustment, effectiveness, and performance. For instance, rationalizers may display higher levels of psychological well-being, enjoy better social reputations, have an easier time influencing their peers, and exhibit superior job performance. Conversely, rationalizers could tend to be unhappy, socially unpopular underperformers, rejected and ineffective due to their self-serving arguments and lack of insight into their own actions. This is analogous to the debate between Taylor and Brown (1988) and Colvin et al. (1995) on the adaptiveness of positive illusions about the self, and it is an empirical question to be addressed in future studies. Some relevant evidence is provided by Uhlmann and Cohen (2005), who find that individuals who rationalize their hiring decisions engage in greater gender discrimination, and yet perceive themselves as more objective and unbiased. This suggests that rationalizations may be associated with favorable subjective self-assessments (see also Dunning et al. 1995), but with suboptimal objective outcomes (for evidence that gender inclusiveness improves group performance, see Hunt et al. 2015; Inglehart & Norris 2003; Woolley et al. 2010). That rationalizers are more likely to make sexist decisions and suffer from an illusion of objectivity would seemingly count as initial evidence against the putative rationality of rationalizations.

Ultimately, the rationalizations-are-rational thesis (per the target article) is important, insightful, and likely to prove generative of further empirical research on attitude-behavior relations, reasoning processes, and human adaptability and performance.

Acknowledgment. This research was supported by an R&D grant from INSEAD.

Heroes of our own story: Self-image and rationalizing in thought experiments

Tomer David Ullman

Department of Psychology, Harvard University, Cambridge, MA, 02138.
tullman@fas.harvard.edu
www.tomerullman.org

doi:10.1017/S0140525X19002280, e51

Abstract

Cushman's rationalization account can be extended to cover another part of his portrayal of representational exchange: thought experiments that lead to conclusions about the self. While Cushman's argument is compelling, a full account of rationalization as adaptive will need to account for the divergence in rationalizing one's actions compared to the actions of others.

Suppose that, like Jason Bourne, you find yourself without memory. In an unfamiliar hotel room, two bodies at your feet, blood splattered on your clothes. In your right hand – a gun. In your left hand – a copy of Cushman's "Rationalization Is Rational." What should you do now?

Cushman argues persuasively that it is entirely reasonable to go about taking actions in the world, in order to recover a policy from actions driven by non-beliefs and non-desires. You could leave your hotel room and start acting and reconstructing, armed with the knowledge that through evolution and habituation your actions probably make sense (also, armed with a gun). But you could also sit down on the hotel bed and have a think.

Thought experiments are part of representational exchange, according to Cushman, a way of learning about the world that is "beyond decision making." There are different accounts of how learning from thought experiments works, with many suggesting an explicit unpacking of mental models that have implicit constraints (see e.g., Clement 2009; Lombrozo 2017; Mach 1897/1976). However, thought experiments often do involve decision making, and the knowledge we gain through them is not necessarily about the world, but about the self. Here I have in mind thought experiments of the more everyday sort, the "would you rather" questions that people like to engage in, as opposed to "what would two blocks tied together to a string do when falling" that only very specific people like to engage in. But many moral reasoning problems fall under this category as well.

Cushman's framework can help explain why such everyday thought experiments are informative, and also why people like to engage in them. Assume that people do not have direct access to their own underlying reasons for action (whether beliefs, desires, habits, or something else) but rather construct a kind of belief-desire theory of themselves (e.g., Gopnik & Meltzoff 1994; Saxe 2009). A thought experiment that asks the thought experimenter what action they would take can engage non-rational (habitual, evolutionarily granted) decision-making modules to produce a hypothetical decision. This decision can in turn be used to update a person's theory of themselves, through a similar mechanism to the inversion of the reward from real actions for others (Baker et al. 2009; 2017). But all of this would be happening without setting foot out of the room. In the same way that Cushman posits an "offline planning" direction from planning to habit in representational exchange, this may be an "offline rationalization."

People take pleasure in answering such thought experiments (McCoy et al. 2019) because information gained in this way is rewarding, in the same way that any information gain or uncertainty reduction may be rewarding in and of itself (Auer et al. 2002). This dynamic of answering from inaccessible modules and updating a theory of those modules can also explain how people can surprise themselves in such thought experiments (McCoy et al. 2019), to the degree that there is a misalignment between the two.

So, Bourne could order some room service, pick up a book, and learn something about himself. However, the overall rational-rationalization account as inverse-policy-learning leaves out a possible central constraint that seems different for inverting one's own policy compared to inverting the policy of another: People are the heroes of their own story. When seeing unconscious cops at his feet, Bourne could reasonably conclude that he's a bad person. Anyone walking into the room at that moment would likely draw that conclusion, so why doesn't Bourne? Fanciful stories aside, there are many situations in which similar behavior driven by similar habits in ourselves and others are rationalized differently, in a way that is skewed in our favor. When I fail to study for a test, it is because the material is not engaging. When you fail to study, it is because you don't like to work hard. In reality, we were both just tired and hungry. Rationalization-is-rational can explain why people try to reconstruct mental variables in these situations as an adaptive behavior. But to the degree that it is adaptive through being often accurate, it seems odd that rationalization would often diverge in this systematic way – unless there was some additional difference to make a difference in this computation. And that difference would itself need to be explained on adaptive grounds.

This is a question about what, if anything, needs to go into the inversion of the policy to make rationalization different for myself and others. It is possible that this is a matter of different input information or missing information in that calculation: In addition to the action of not studying, I am also privy to certain mental states like the fact that I am not lazy. But this seems to be begging the question; the whole point of rationalization is that it reconstructs such states where there were none, without the awareness of the person doing the rationalization. An alternative is that there is an overarching, adaptive principle that ensures that rationalization for one's self is more in one's favor than inverse planning for others. This would be akin to one's own beliefs and desires being in line with what one sees as good or desirable. But if this comes at the expense of accurate inference, why have it at all?

In short, I am on board with Cushman's account of rationalization as adaptive, and as part of a broader account of representational exchange. If anything, I think this account can be broadened to capture the engaging aspect of thought experiments that involve making a decision. However, a full account of the functional role of rationalization will need to account not only for its self-benefitting nature, but also its self-serving one.

The social function of rationalization: An identity perspective

Jay J. Van Bavel, Anni Sternisko, Elizabeth Harris and Claire Robertson

Department of Psychology, New York University, New York, NY 10003.
jay.vanbavel@nyu.edu as10039@nyu.edu
eah561@nyu.edu cer493@nyu.edu
https://as.nyu.edu/content/nyu-as/as/faculty/jay-van-bavel.html

doi:10.1017/S0140525X19002097, e52

Abstract

In this commentary, we offer an additional function of rationalization. Namely, in certain social contexts, the proximal and ultimate function of beliefs and desires is social inclusion. In such contexts, rationalization often facilitates distortion of rather than approximation to truth. Understanding the role of social identity is not only timely and important, but also critical to fully understand the function(s) of rationalization.

In "Rationalization Is Rational," Cushman offers a serious consideration of the function of rationalization that is long overdue and represents an important contribution to the literature on human psychology. He states that "the ultimate function of belief is to represent true properties of the world, and the ultimate function of desire is to represent the fitness consequences of these properties." (target article sect. 4.4.1, para. 13). In this commentary, we offer an additional function of rationalization. Namely, in certain social contexts, the proximal and ultimate function of beliefs and desires is social inclusion. In such contexts, rationalization often facilitates distortion of rather than approximation to truth. Understanding social identity is not only timely and important, but also critical to fully understand the function(s) of rationalization.

Although Cushman's discussion of rationalization focuses on the self, we think he understates the truly social nature of the self. In our view, any self-related process, including rationalization, must take into consideration how people categorize themselves in a social context (Turner et al. 1994). Thus, the functions of rationalization need to be understood within the context of intragroup and intergroup dynamics.

We argue that rationalization of actions and beliefs of fellow group members, and of one's own actions and beliefs in relation to these group members, can enhance the fitness of an individual by maintaining one's position in the group, even if the rationalization leads to false beliefs about the world. Rationalization and adoption of group-based beliefs can also help fulfill more proximal goals, including the need to belong, obtain status, understand the social world, and feel morally justified (Van Bavel & Pereira 2018).

The domain of politics offers extensive evidence that political identities motivate people to resist factual evidence that undercuts their group affiliation, to rationalize lies from group leaders, to believe identity-bolstering fake news, and to generate politicized conspiracy theories. These examples highlight the critical role of social identity in rationalization:

- *Resist evidence.* People often discount or rationalize evidence that contradicts their firmly held political beliefs or party affiliation. For instance, people are less likely to update their political views in the face of counterevidence compared to their non-political views (Kaplan et al. 2016). This belief resistance was associated with activity in the prefrontal cortex – suggesting a role for motivated reasoning or rationalization. In some cases, exposure to opinions from political out-group members can even backfire – making people more entrenched in their political beliefs than before (Bail et al. 2018).
- *Rationalize lies.* Cushman argues that rationalization allows people to translate gut instincts into rational thoughts that "represent[s] true properties of the world." However, people readily rationalize false information when it is propagated by party elites and aligns with their political identity For example, all Trump-branded hats are made in the United States, but when Clinton supporters were told to imagine that Trump would make his merchandise outside the United States if it were cheaper to do so, they felt it would be less unethical to lie that this merchandise *was* made outside the United States, and that political elites who espouse these lies deserve less punishment (Effron 2018).
- *Believe fake news:* People also rationalize fake news that is positive about one's in-group or negative about one's out-group. For example, Democrats were more likely to believe negative fake news about Republican politicians than negative fake news about Democratic politicians, and vice versa for Republicans (Pereira et al. 2019). People are typically motivated to hold true beliefs, but in the case of identity-bolstering fake news, it can be more beneficial to rationalize these false beliefs as true. In believing them, people can share similar beliefs with in-group members and maintain positive beliefs about the group.
- *Generate conspiracy theories.* Conspiracy theories connect different, unrelated, and inconsistent events in a way that seems meaningful and rational. As such, conspiracy theories can help uphold a positive group-identity under the guise of rationality. For instance, some scholars argue that *conspiracy theories are for losers,* such that the loss of political power increases conspiracy theory beliefs (Uscinski & Parent 2014). Indeed, prior to the 2012 U.S. presidential election, Republicans and Democrats were similarly likely to expect electoral fraud. However, after President Obama was re-elected, Republicans were more likely to believe that electoral fraud had occurred (Edelson et al. 2017). Reducing the their loss to a conspiracy allowed Republicans to rationalize and uphold positive beliefs about their in-group.

This is a sample from a large literature exploring the social function of rationalization. Factors that increase identification with political parties or movements can increase the value of rationalization given that it may help people remain in good standing with fellow group members (see Van Bavel & Pereira 2018). Furthermore, rationalizing actions of in-group elites can reduce accountability for harmful behavior and create conflict with out-group members, thus increasing polarization. At the same time, polarization can increase commitment and identification with one's in-group, thereby motivating rationalization. Thus, aspects of the intergroup context, like polarization, can both amplify rationalization and result from group-based rationalization.

Understanding the role of social identity in rationalization is not only critical for understanding the function(s) of this psychological process, but also clarifying when and why features of the context will elicit and result from rationalization. For instance, situations that increase the salience of identities or the norms associated with those identities will impact rationalization. These forms of rationalization not only help an individual maintain or increase their standing within the group (which can promote well-being and survival), but also ensure that the group maintains cohesion during intergroup competition.

The rationale of rationalization

Walter Veit[a], Joe Dewhurst[b], Krzysztof Dołęga[c], Max Jones[a], Shaun Stanley[a], Keith Frankish[d] and Daniel C. Dennett[e]

[a]Department of Philosophy, University of Bristol, Bristol, BS8 1TH, United Kingdom; [b]MCMP, LMU Munich, 80539 Munich, Germany; [c]Institut für Philosophie 2, Ruhr-University Bochum, Universitätsstraße 150, 44801 Bochum; [d]Department of Philosophy, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom and [e]Center for Cognitive Studies, Tufts University, Medford, MA 02155.

wrwveit@gmail.com https://walterveit.com/
joseph.e.dewhurst@gmail.com https://joedewhurst.weebly.com/
krzysztof.dolega@rub.de https://www.krysdolega.xyz/
Max.Jones@bristol.ac.uk http://www.maxjonesphilosophy.com/
shaun.stanley@bristol.ac.uk8
k.frankish@sheffield.ac.uk https://www.keithfrankish.com/
Daniel.Dennett@tufts.edu https://ase.tufts.edu/cogstud/dennett/

doi:10.1017/S0140525X19002164, e53

Abstract

While we agree in broad strokes with the characterisation of rationalization as a "useful fiction," we think that Fiery Cushman's claim remains ambiguous in two crucial respects: (1) the reality of beliefs and desires, that is, the fictional status of folk-psychological entities and (2) the degree to which they should be understood as useful. Our aim is to clarify both points and explicate the rationale of rationalization.

Post hoc rationalization, that is, retrospectively attributing or constructing "hidden" beliefs and desires inferred from how one has behaved in the past, has traditionally been seen to threaten the idea that humans are "rational," since it happens subsequent to the process under consideration. If the relevant mental states that are supposed to rationalise an action only come into existence after the action has occurred, then they cannot be treated as the cause of that action. However, Cushman argues that a post hoc process of this kind can still be seen as "rational" in the sense that it constructs new beliefs and desires that both serve a useful function and track some underlying adaptive rationales that have shaped the behaviour being rationalised. Rationalization, according to Cushman, is supposed to be a "useful fiction." We think that this proposal invites two serious ambiguities: first, to do with the ontological status of the mental states that are the outputs of rationalization (i.e., folk-psychological states like beliefs and desires) and, second, to do with the degree to which they should be understood as useful and representative. We will address each ambiguity in turn, using our resolution of the latter to help resolve the former.

Throughout his article, Cushman seems to assume a fairly robust understanding of what beliefs and desires are, framing them as functionally discrete internal states with determinate contents. He is committed to the idea that there is a crucial distinction between "real" reasoning processes, which involve operations on beliefs and desires, and the fictional ones produced by rationalization, which don't involve any such operations. Rationalization, on his account, seems to play the role of a process of self-interpretation in which one authors fictions about the causes of one's own behaviour. Drawing these distinctions might not be as easy as Cushman suggests, if there is no principled "dividing line between *genuine* belief-talk or agent-talk and mere *as if* belief-talk and agent-talk" (Dennett 2011, p. 481). Indeed, the lack of such a dividing line similarly arises for agential descriptions or "rationalizations" in evolutionary biology (see Dennett 2019; Okasha 2018; Tarnita 2017; Veit 2019). Without such a dividing line, however, it is unclear what the ontological status of beliefs and desires is supposed to be. If Cushman were to deny that there are anything at all like beliefs and desires prior to the rationalization process, making the folk-psychological states produced by this process entirely fictional, he would fall close to eliminative materialists such as Paul and Patricia Churchland (Churchland 1981; 1986). We do not think that Cushman would like to endorse this option, as he seems quite committed to the existence of beliefs and desires. The other option, then, and this is a move we recommend for Cushman, is to commit to the existence of some sort of proto-mental states prior to the rationalization process, in which case we think it is unclear in what sense the output of the rationalization process also constitute fictional entities. Of course, the rationalization process might influence or replace these proto-mental states via a narrative process that we could call fictional, but it is no longer the mental states themselves that are fictions, rather the process that produces them.

This brings us to the second ambiguity: In what sense can fictional mental states (or processes) be understood as useful? Cushman clarifies that these fictions can be useful even when they are not "perfectly accurate representations" by appealing to Dennett's (1987) "intentional stance," according to which the attribution of beliefs and desires are understood as nothing more than a way of tracking observable patterns in behaviour (or the categorical bases of those patterns) and have no further ontological status *inside* the system. However, this comparison reveals a tension in his dual conception of folk-psychological states. Dennett's intentional stance assumes that habit, instinct, norms, and so on, may all support rational patterns of behaviour, and that this is all that is needed for a system to manifest genuine beliefs and desires. It is true that these processes support rational responses that make it worth extracting information from them via rationalization (i.e., by adopting the intentional stance) and then re-presenting this information in a rich belief/desire format. Reformatted in this way, beliefs and desires take the form of the linguistic utterances that Dennett (1987) originally called "opinions" and Frankish (2004) has more recently called "superbeliefs." For us, richness is a matter of having a discrete representational vehicle, such as that provided by natural language, but it is not clear that this is what Cushman has in mind when he talks about beliefs and desires.

As we see it, there are two broad ways to achieve such a rich conception of belief, either internal or external. On the internal conception, that is, traditional computationalism, this vehicle is a neural one, and beliefs are formed and processed at a subpersonal level. On the external conception, the vehicle is natural language, and beliefs are formed and manipulated at a personal level by agents themselves, as a way of describing and regulating their own and others' behaviour. Forming a rich belief, that is, an opinion or superbelief, is like adopting a policy or making a bet on truth – we commit to taking a sentence as an expression of truth and regulate our other utterances and commitments accordingly. Cushman seems to espouse a version of the former interpretation, but we think that the latter interpretation is to be preferred, as it can help to resolve the two ambiguities outlined above.

Once this external approach is adopted, the sense in which rationalization is *fictional* becomes clear: It involves the construction of a narrative that is strictly false with regard to the underlying mechanisms, but nonetheless captures real patterns in the behaviour generated by those mechanisms. We propose to interpret rationalization as the process of taking the austere "proto-beliefs" manifested in behaviour and transforming them into superbeliefs or opinions (i.e., rich, linguistically formatted beliefs and desires) via the application of the intentional stance to one's own behaviour. Taking this can help to resolve the ambiguities described above, provided that Cushman is willing to adopt this distinction between the austere beliefs that are implicit in all (seemingly) intelligent behaviour, and the explicit, linguistically mediated beliefs that are the outcome of the rationalization process. The latter could be seen as fictional, in the sense that they only came about as the result of a story that we tell about our own behaviour, and yet they are also real, in the sense that they do accurately capture (and help to track) our behaviour (even if they do not accurately describe the processes underlying that behaviour). By coming to be explicitly represented in natural language, expressing normative commitments, they can also indirectly influence our future behaviour. In short, we think rationalization should be treated as the reverse engineering of what Dennett (2017) has called "free-floating rationales," that is, instinctive behavioural patterns, like avoiding snakes or heights, that are not explicitly encoded but nonetheless make rational sense. Similarly, the underlying *reasons* that are implicit in our behaviour can be inferred (or rather uncovered) via rationalizations, which can then lead to further behavioural improvements by engaging in explicit rational deliberation. This is the rationale of rationalization.

Hard domains, biased rationalizations, and unanswered empirical questions

Stephen E. Weinberg[a] and Jonathan M. Weinberg[b,1]

[a]Department of Public Administration, Rockefeller College of Public Affairs and Policy, University at Albany, State University of New York, Albany, NY 12222 and [b]Department of Philosophy, University of Arizona, Tucson, AZ 85721-0027.
sweinberg@albany.edu jmweinberg@email.arizona.edu
https://www.albany.edu/rockefeller/faculty/stephen-weinberg

doi:10.1017/S0140525X1900222X, e54

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Abstract

Cushman raises the intriguing possibility that rationalization accesses/constructs intuitions that are not otherwise cognitively available. However, he substantially over-reaches in arguing that rationalization is mostly right on average, based on claims that the process must have emerged adaptively. The adaptiveness of "bounded rationalization" is domain specific and is unlikely to be adaptive in a large number of important applications.

Cushman argues that rationalization is rational because it lets us construct useful fictions, which on average will yield "true beliefs and adaptive desires" (sect. 3.5, para. 5). He presupposes that the benefit of rationalized beliefs will offset any such costs incurred when false beliefs are produced. But how often may we expect the benefit-to-cost ratio to go the right way? It seems to us to be an empirical question that Cushman perhaps does not adequately address. (And note that the deck starts stacked against him, because rationalization will always involve at least some false beliefs about the recent past states of one's own mind.) We will stipulate that it is prima facie reasonable to assert that cognitive processes are adaptive within their evolutionarily relevant domains. Cushman extends this domain beyond the Paleolithic era by bringing in the idea of cultural evolution (which puts us on a time scale of decades), but in an age of Google, algorithmic stock market trading, and six-tier pension plans, it is unreasonable to expect culture to adapt in time to technological changes.

That leaves the process of reinforcement learning as an adaptive mechanism that might calibrate my rationalization process into something useful. For high-frequency, feedback-rich tasks like driving to work, I might be expected to recalibrate, eventually changing my intuitions if I move from Boston to Brisbane. For complicated, low-repetition, poor-feedback tasks like saving for retirement, I can at best assert a form of bounded rationalization, subject to all the same cognitive challenges as any other information-processing task (for a recent treatment, see Benjamin 2018).

For example, myside biases might manifest as overconfidence in my ability to pick stocks – and to retain that overconfidence after years of failing to beat the market (see Daniel & Hirshleifer 2015, or see our father's investment returns for the last 40 years). In general, in the presence of confirmatory biases, even large amounts of feedback may well fail to drive people to correct beliefs (see, e.g., Rabin & Schrag 1999).

It gets worse. As Cushman himself notes, our theory of mind capacities are "biased to perceive all behavior as rational, even though much behavior is not" (sect. 3.5, para. 6). But he has unfortunately failed to follow up on a consequence of this psychological fact: This bias will also afflict the process of rationalization. Rationalization will thus only entertain a radically restricted subset of possible hypotheses, and accordingly, whenever the relevant truths about the environment fall outside that subset, rationalization will not be beneficial.

Cushman's diagram in Fig. 2b presents three unconscious systems: instincts, norms, and habits. He argues that these systems join with reasoning to influence action, and those actions then provide signals to the underlying systems. However, this list of unconscious processes leaves out such possible *intuitively irrational* influences as myside biases, present-biased discounting, primacy and recency effects, and so forth. Because such biases will not be perceived as rational, rationalization will both fail to extract any information value from them where they have such value and make it harder for agents to correct for them when they lack such value.

Consider the example of saving for retirement. If I observe that older workers are relying heavily on their defined benefit pensions, and that fact unconsciously becomes the basis for rationalizing skimping on my own retirement contributions, then I will have made a very costly error. Even if I have a defined benefit pension plan, it is very likely to be stingier than that of the older workers whose choices have informed my rationalization. If I knew that my rationalization was really based on the older workers, it would be easier to notice that they have a different plan than I do.

Cushman argues that rationalization will "occasionally be maladaptive" (sect. 1.1, para. 10). We believe this statement is highly optimistic. We believe that modern society offers a large number of domains that are far removed from the conditions in which our biology and culture evolved, and which are too dynamic or feedback-poor to have calibrated reinforcement learning (see Shanteau 1992). These domains include high-cost, low-frequency decisions like what education to get, how to save for retirement, and whether/whom to marry – none of them choices that provide for much learning, but which are extremely consequential. How often will people be calibrated enough for rationalization to be reasonably adaptive? What are the relative costs of adaptive versus non-adaptive rationalizations? Without positive evidence to resolve such questions, rationalization cannot yet be viewed as rational.

However, we would conclude by emphasizing that we view this as an open empirical question. One possibility Cushman may wish to explore, in trying to arrive at a positive verdict, is that some of these biases will not merely corrupt rationalization, but also be compensated by it. Consider our example of being an overconfident investor. If I was otherwise perfectly adapted to the modern stock market, then a misbelief could immediately push me into bad choices. However, behavioral scientists have documented entire hosts of "non-rational influences." For example, Benartzi and Thaler (1995) demonstrate that a different influence, myopic loss aversion, can paralyze investors. These countervailing biases, one promoting investment and one suppressing it, could lead someone to better investment decisions than they would make if subject to only one of them. See Bahaddin et al. (2019) for further discussion of countervailing biases and Akerlof and Yellen (1985) for the implications of starting far away from the optimum. The presence of multiple biases supports Cushman's claim that even inaccurate rationalizations could potentially be adaptive.

We find Cushman's descriptive claim that rationalization can be viewed as a form of representational exchange to be a fruitful one, and we applaud his opening up the question of whether rationalization has perhaps gotten an unfairly bad rap. We contend that his further normative claim on behalf of rationalization unfortunately falls short. We do not, at this time, have reason to think that rationalization is rational, at least, not based on Cushman's arguments.

Note

[1] Authors are listed in the order of "Mom always liked you best."

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Author's Response

Rationalization as representational exchange: Scope and mechanism

Fiery Cushman

Department of Psychology, Harvard University, Cambridge, MA 02138
cushman@fas.harvard.edu https://cushmanlab.fas.harvard.edu

doi:10.1017/S0140525X19003261, e55

Abstract

The commentaries suggest many important improvements to the target article. They clearly distinguish two varieties of rationalization – the traditional "motivated reasoning" model, and the proposed representational exchange model – and show that they have distinct functions and consequences. They describe how representational exchange occurs not only by post hoc rationalization but also by ex ante rationalization and other more dynamic processes. They argue that the social benefits of representational exchange are at least as important as its direct personal benefits. Finally, they construe our search for meaning, purpose, and narrative – both individually and collectively – as a variety of representational exchange. The result is a theory of rationalization as representational exchange both wider in scope and better defined in mechanism.

R1. Introduction

It is a delight to receive such thoughtful commentaries, and a gift to be able to reply. They offer many important improvements to the target article. My goal is to make these as clear as possible. To this end, several sections of this reply summarize the most valuable revisions to the theory of rationalization as representational exchange offered by the commentaries (see R2.1, R4.3, R5.4, and R7.4). The target article focused the representational exchange function of rationalization, but the commentaries show that is important to also acknowledge its self-serving functions and to keep these distinct. The target article focused on post hoc rationalization, but these commentaries extend it to cover antecedent and more dynamic cognitive processes. The target article focused mostly on the personal benefits of representational exchange, but these commentaries extend it to cover the many important interpersonal and collective benefits. Most significantly, several of these commentaries translate the academese of the target article ("rationalization extracts implicit information") into the native language of human experience. We constantly try to find, or make, meaning in our lives. We do this individually, and we do it collectively. We do it through self-reflection: By interrogating the purpose of our actions and attitudes and by striving for consistency and coherence among our beliefs. And, whether we realize it or not, we do it for self-improvement. The theory of representational exchange offers a natural account of these behaviors. I conclude by considering the challenges, opportunities, and pitfalls of self-improvement by rationalization.

Perhaps, then, I should not have titled the target article "Rationalization Is Rational," but instead "Self-Reflection Is Useful." Although I can't shake the feeling that there was something right about the original title, I'll have to think a little harder about what that might be.

R2. Motivated reasoning versus representational exchange

There must be at least two kinds of rationalization. One of these is proposed by the target article, and **Ullman** calls it "self-benefitting." Through representational exchange it improves our beliefs, desires, and ultimately our behavior. But several commentaries urge that we not overlook the other kind of rationalization, which Ullman calls "self-serving" (see also **Quilty-Dunn**), and **Ellis & Schwitzgebel** call "rationalization in the pejorative sense." It is basically a variety of motivated reasoning. In order to convince ourselves or others that we did the right thing, we invent explanations for our behavior that cast it in favorable light. Such rationalizations do not distill a liquor from the mash of non-rational influences, but instead slap a fraudulent label on a jug of rotgut. Ellis & Schwitzgebel, Quilty-Dunn, **Tierney & Uhlmann**, **Altay & Mercier**, **D'Cruz**, and **Brody & Costa** argue that most rationalization is self-serving. Whether or not it is more common, it is certainly more prototypical. Perhaps Ellis & Schwitzgebel say it best: "Rationalization has a bad name for good reason, and that will be missed on Cushman's theory."

How badly is it missed? The target article certainly doesn't deny that rationalization is sometimes self-serving. To the contrary: "This family of theories likely explains a part of the function of rationalization. If the present account also explains a part, then it is a complementary but largely independent explanation" (target article, sect. 2.3, para. 4). Yet, by calling these two kinds of rationalization by the same name, the target article implied an unnecessary competition between them.

The wiser course, urged by **Ellis & Schwitzgebel**, **Ullman**, **Quilty-Dunn**, and **D'Cruz** is to allow each its own name. Following Ullman, I will call them self-serving (i.e., rationalization as motivated reasoning) and self-benefitting (i.e., rationalization as representational exchange). Self-benefitting rationalization is not necessarily distinct from self-serving rationalization because of consequences – that is, adaptive versus maladaptive outcomes. Presumably both are usually good for people but sometimes bad. Neither is it necessarily distinct in its mechanisms. It is possible that both forms of rationalization share a common mechanism. Rather, what distinguishes self-benefitting rationalization from self-serving rationalization is its function: representational exchange. Following from this function, self-benefitting rationalization should generally improve the quality of reasoning, while self-serving rationalization will generally diminish it.

Of course, the key question is whether self-benefitting rationalization actually happens. Insofar as the human mind is designed to rationalize, is representational exchange ever the ultimate goal? Several commentaries raise important doubts. For instance, existing research shows that rationalization is often empirically associated with a worse, not better, quality of reasoning and action. Summarizing some this evidence, **Tierney & Uhlmann** entertain the possibility that "rationalizers could tend to be unhappy, socially unpopular underperformers, rejected and ineffective due to their self-serving arguments and lack of insight into their own actions" – a vision more or less shared by **D'Cruz**, **Brody & Costa**, and Tierney & Uhlmann. (**Quilty-Dunn** shares the view that rationalization degrades reason, but views it as a useful "drive state that pushes us to reorient our beliefs in ways that keep our self-esteem up, even when we are irrational, unstable, or immoral, and thereby keep ourselves motivated by avoiding the awful truth.")

In support of these claims, **Tierney & Uhlmann** review evidence the rationalization of gender bias in hiring decisions is associated with greater gender bias in hiring decision. **Brody & Costa** review the evidence that rationalization of violence and criminality is associated with greater antisocial behavior. **Quilty-Dunn** reviews evidence that rationalization of meat consumption is associated with the belief that it is "necessary, normal, natural, and nice."

If you believe that gender bias, violence, criminality, and eating meat are bad things, then surely it is bad to rationalize them. But does that mean that rationalization itself is bad? Obviously, good objects can be used for bad ends: A chef's knife makes a fist fight worse, but it is awfully handy for cooking. When we look for bad effects of rationalization, we can find them; what would happen if we instead looked for good effects? For instance, is rationalizing charitable acts associated with more prosocial behavior?

Brody & Costa review evidence that high levels of rationalization characterize several mental disorders. They conclude that rationalization is generally maladaptive. But it is easy to see the flaw in arguments of this kind. Anxiety disorders are characterized by excessive fear, but this doesn't mean that fear is generally maladaptive. Presumably many mental disorders arise from the dysregulation of thought processes that, in moderation, are essential to proper psychological function. In fact, as **Saxe & Nettle** note, self-benefitting rationalization (i.e., critical self-reflection about the causes of one's behavior) is essential to cognitive behavioral therapy, one of the most effective, evidence-based treatments available for mental disorders. It is an advantage of the representational exchange framework that we can identify both why this kind of behavior can be helpful and also how its misapplication can cause harm.

De Neys summarizes further evidence consistent with self-benefitting rationalizations. Across several experiments, they find that rationalization is often used in service of finding the right explanations for the right actions, but in cases where the actions themselves were a product of intuition. For instance, consider the famous "bat and a ball" question that appears in the Cognitive Reflection Task (Frederick 2005). When people report the correct answer to this word problem, they often cannot say exactly why it is correct. Rather, they have to deliberate for a moment about why it is correct, eventually explaining their answer in terms of sound reasons, but ones that were only rendered explicit by post hoc rationalization. This illustrates rationalization of a self-benefitting kind.

R2.1. Revision: Distinguishing two functions of rationalization

The pejorative model of rationalization currently dominates. It proposes that while rationalization may have some benefits, improving reasoning is not among them. Not surprisingly, experiments designed to show examples of rationalization in the pejorative sense tend to find that it is associated with bad outcomes. The target article was not intended to deny the existence of self-serving rationalization. It exists. We've documented it. We understand it pretty well. Rather, the contribution of the main article is to point out that the same mechanism – or a very similar one – can also improve reasoning.

Consider an analogy to stereotyping. Prototypically, stereotypes are bad things that make us worse off. This occurs in part because stereotypes can become a vehicle for motivated reasoning, allowing us to distort our factual understanding of the world in order to serve our social interests. Yet, everybody also recognizes that there is also a "rational" side to stereotypes. Given that our minds must make the best use of limited data and limited computation, it makes sense to understand and predict the behavior of tokens (e.g., individual people) by drawing on statistical generalizations at the level of the type (e.g., social group). A common form of computation and representation at the mechanistic level – the stereotype – can serve two quite different functions. One is generally helpful; another can be extraordinarily harmful.

A key outstanding question is whether, and to what extent, self-serving and self-benefitting rationalization share a common mechanism. There is a spectrum of possibilities. At one extreme, it may be that fully distinct mechanisms engage in rationalization of self-serving and self-benefitting kinds, and that the only connection between these mechanisms is in the abstract form they take. On this view, they are as distinct as the umbrella and the parasol sitting beside each other in a closet – although similarly constructed, they are both physically and functionally distinct. Or, it may be that there is a single, common mechanism that happens to serve two independent purposes. On this view, rationalization is like a roof: A single object that serves the dual purposes of sheltering from rain and from the sun.

R3. Does rationalization really happen?

Tierney & Uhlmann, **Stafford**, **Dahl & Waltzer**, and **Pärnamets, Johansson, & Hall** (**Pärnamets et al.**) argue that rationalization is a rare and weak force in our minds. These concerns must be taken seriously because they are supported by some of the very evidence cited in the target article. In a study by Sharot et al. (2010), for instance, the experimentally induced perception of having chosen one vacation destination over another brought about only a 0.06 point shift in subsequent preference, on an 6-point scale.[1] Although there are some published large effect sizes of rationalization, these commentaries make a persuasive case that the effects of standard laboratory methods (such as the "free choice paradigm") are usually quite small.

It is notable, however, that standard experimental methods of eliciting rationalization focus specifically on cases where it would be *irrational*. In other words, social psychologists tend not to induce rationalization by prompting people to acts in an adaptive manner and then, through a bit of self-reflection, drawing useful lessons from it. Instead, they tend to devise circumstances in which the experimenter has cleverly induced somebody to act in a random or even maladaptive manner, and thus any rationalization would necessarily corrupt reasoning.

This is a useful approach from the standpoint of experimental design. In order to draw strong inferences about psychological mechanisms, it is helpful to put a participant's behavior under experimental control and then demonstrate that they do something predicted by the mechanism, but otherwise irrational and, therefore, very difficult to explain in any other way (see Saxe 2005).

Unfortunately, the very same features of an experiment that allow us to draw strong inferences may tend to generate weak effects (Mook 1983). If rationalization "approximate(s) a well-understood form of rational inference" (target article, sect. 3.3, para. 3) then, when people are tricked into believing they chose Greece over Thailand, they may rationally infer little from it. Rather, they may conclude that this "choice" was mostly a matter of chance (Gershman 2019; Hawthorne-Madell & Goodman 2019). And, therefore, they may exhibit little change in their beliefs or attitudes.

To see why, it helps to focus on the details of this particular experiment: Although exquisitely designed for strong causal inference, it is wholly unlike ordinary life. Participants are instructed to click on one of two nonsense words that, they are told, contain subliminal information eliciting an unconscious decision between vacation destinations. In fact, the experimenter randomly assigns the participant one vacation destination or the other. Upon "learning" that their "choice" was Greece (for instance), the participant might begin to introspect a bit. "How do I feel about Greece? Do I really have strong feelings that it is a better vacation destination?" Due to the very logic of this experiment, on average the answer to this question must be "no," since the participant's apparent choice is randomly assigned. In light of the peculiar circumstances of the choice, and lacking any actual underlying attitude supporting it, people may reasonably conclude that there are few "true beliefs" or "useful desires" to be gleaned. Evidence for rationalization in this context allows for a strong causal inference, but we ought to expect the effect size to be very small.

Contrast this with a person who is *actually* choosing between vacations in Greece and Thailand – say, in a travel agency, with colorful brochures and detailed information. They feel a strong, intuitive pull toward Greece, but can't yet put their finger on why. They try to explain their feelings to themselves and to the travel agent. In this circumstance, they might reasonably conclude that there is something true, and something important, to be learned during self-reflection.

Dahl & Waltzer raise a second objection to the thesis that rationalization makes much contribution to human thought and behavior. They write, "If rationalization were widespread, we would have to abandon a premise of much discourse: that our beliefs and actions are generally based on reasons." Along similar lines, **Tierney & Uhlmann** summarize findings that "Time 1 explicit attitudes predict Time 2 behaviors far better than past behaviors predict future self-reported attitudes, calling into question the prevalence of post hoc rationalizations for past actions… [and] relegating the 'rationalizations are rational thesis' to address only a small portion of the attitude-behavior relationship."

These objections depend on the premise that rationalization is opposed to reason, and so the more influential one is, the less the other must be. Yet, the very premise of the target article is that rationalization can actually improve reasoning. Indeed, the theory depends upon the premise that we are often rational creatures whose Time 1 beliefs and desires appropriately shape our Time 2 behaviors; after all, there is no point exchanging representations in order to improve reasoning if it is impotent or inert! (See sect. R5, where the target article is defended against the critique that we never actually reason).

By analogy, consider the proposition: "Humans frequently sleep in order to benefit their waking lives." It hardly counts as an objection to say, "But if we sleep so frequently, this leaves no time for us to be awake!" Likewise, "Research suggests we accomplish nearly all tasks while awake, relegating to sleep to an inconsequential role!" Just as a few hours of sleep render our many waking hours more productive, so too might a few moments of rationalization support many moments of reasoning.

Berthelette & Kalbach agree that nearly all adaptive behavior traces back to reasoning, but they arrive there by a different route. They argue that when habits work well – that is, when they are appropriately attuned to circumstances and recommend the correct behavioral policy – they are "best explained as being caused by a standing belief and desire" Thus, there can be no rationalization "because the relevant beliefs and desires were there all along." The implied psychological model seems to be that habits arise exclusively through the stamping-in of rational actions through repetition (something like practicing a piano piece until it becomes "muscle memory"). This kind of habit formation is well documented (Dezfouli & Balleine 2012; 2013; Dezfouli et al. 2014; Miller et al. 2019). But it is also well-documented that we cache the context-dependent value of actions based on a history of reward (Schultz et al. 1997; Glimcher 2011; Morris & Cushman 2019). In such cases the habitual action does not depend, either presently or historically, on a generative causal model – that is, on beliefs (Dolan & Dayan 2013). In other words, habits are not always just "precompiled rational acts." Rather, they are sometimes formed by a computationally and representationally distinct process.

In summary, the objection that "rationalization doesn't happen" begins with the premise that most rationalization is self-serving. Standard social psychology paradigms are well designed to estimate the effect of this "bad" kind of rationalization, and to the extent that it occurs, it stands in opposition to sound reasoning. In contrast, however, the target article proposes a second kind of rationalization. Its influence may be systematically underestimated by standard paradigms, and it is not opposed to reason. Future experiments should explore whether, how often, and how powerfully it shapes our thinking.

R4. The mechanics of rationalization: Prioritized representational exchange

Several commentaries focused on the mechanisms underlying rationalization and representational exchange (**Harmon-Jones & Harmon-Jones**, **Sievers**, **Simon & Holyoak**). The target article focused on a simple case: The post hoc rationalization of action. Here, a person has already performed an action and adjusted their beliefs and desires in a manner that would have rendered the action rational.

The commentaries offer two basic extensions of this model. First, rationalization is often antecedent: "we rationalize our actions not only after we perform them, but also before we perform them and sometimes as a condition of performing them" (**Ellis & Schwitzgebel**). Similarly, we often rationalize not an action, but an attitude or emotion (**Railton**). Second, rationalization is not merely a process of updating beliefs and desires conditional on action, but a more general process of achieve coherence among a variety of mental state representations, with multidirectional influences.

R4.1. Antecedent rationalization

Sievers and **Simon & Holyoak** assemble impressive evidence for antecedent rationalization, in which a person adjusts their beliefs and desires prior to action (see also **D'Cruz**, **Ellis & Schwitzgebel**, and **Tierney & Uhlmann**). As Sievers points out, "Nothing entailed by antecedent rationalization precludes the correctness of Cushman's model of rationalization as a representational exchange mechanism. It does, however, demand an expansion of that model." This is an important point: Antecedent rationalization happens. It fosters representational exchange when the anticipated action is influenced by a non-rational adaptive control mechanism (e.g., habit, norm, or instinct).

In a paradigmatic case of antecedent rationalization, we bring beliefs and desires into alignment with a non-rational system, such as habit. For instance, suppose that cached value representations (habit) recommend choosing cake instead of tea for dessert, but that a planning system (reason) recommends tea over cake. We would adjust our reasons for tea ("I've already eaten too

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

many calories in this meal…") so that they become aligned with our habitually favored choice ("…but it's Lisa's birthday, and she'd be disappointed if I didn't have some cake"). In this case our cached values remain constant while our reasons change.

Extending this idea, **Railton** proposes that we often rationalize our emotions, and not just our actions. Thus, for instance, suppose that a coworker inspires intense feelings – perhaps admiration, jealousy, anger, love or contempt. Even if these feelings do not move us to any particular action, they still may shape our beliefs by rationalization. In this case, we are learning what to think or want not by observing how we act, but instead by observing how we feel. In this sense it is akin to antecedent rationalization, which is also prompted by a thought, rather than an act.

But once we acknowledge that certain non-rational forms of thought (e.g., emotion, or habit) can antecedently adjust the representations that contribute to reasoning (beliefs and desires), surely we must also acknowledge that the influence could also run the opposite way, as well. To return to the example above, couldn't the reasons remain the same, while the habit-based values are adjusted so that tea, rather than cake, becomes the habitual response (e.g., Gershman et al. 2014)? These are both instances of representational exchange, but they push information in opposite directions. In one case, information flows from habit to reason; in the other case, from reason to habit.

Presumably, the controller that you "trust" more – the one that is most likely to generate the correct policy – should both determine your action and serve as the source for representational exchange. Deciding which controller to trust is often called the *metacontrol* problem. Classically, metacontrol is construed as a process of deciding which system governs action. (In this manner, it is pivotal to post hoc rationalization, which depends upon the control system underlying action). Importantly, however, it could also be repurposed as a mechanism for determining the direction of representational exchange, for instance, during antecedent rationalization. Metacontrol determines which system should be trusted, and it is an active area of research with several promising theoretical approaches in development (e.g., Daw et al. 2005; Griffiths et al. 2015; Kool et al. 2017; Shenhav et al. 2013).

In summary, antecedent rationalization surely occurs. But it draws our attention to a crucial problem faced by the theory of representational exchange: How is the direction of exchange established? The existing literature on metacontrol, which is similarly concerned with the problem of deciding which control systems to trust, offers promising leads. It also immediately suggests a more dynamic model in which rational and non-rational processes continually exert influence upon each other in a manner prioritized by metacontrol.

R4.2. Coherence

Simon & Holyoak and **Harmon-Jones & Harmon-Jones** both contrast the theory of representational exchange with alternative models focused on coherence. To quote Simon & Holyoak, "In models of *coherence-based reasoning*… complex decision situations are represented by networks in which the relevant variables are interconnected via excitatory and inhibitory links. Constraint satisfaction mechanisms settle the network into a stable state of coherence, in which mutually supportive connections (i.e., those that "go together") activate one another and collectively inhibit their rivals" (emphasis added).

Simon & Holyoak's commentary has a lot to say about the mechanisms by which coherence is achieved, but little to say about why coherence itself is a useful property of the mind. Several other commentaries furnish possible adaptive explanations. **Harmon-Jones & Harmon-Jones** appeal to the idea that coherence reduces internal conflict: "The organism will have a difficult time enacting a course of behavior when trying to follow the action tendencies evoked by conflicting cognitions." And we have already summarized several related proposals that coherence has the useful social function of rendering our behavior more predictable to others (**Levy**; **Saxe & Nettle**; **Stanford**, **Thomas**, **& Sarnecka** [**Stanford et al.**]). Notably, however, these explanations focus on intrinsic benefits of coherence: that is, the lack of discrepancy or conflict between representations. Rather than positing that we prioritize "better" representations and adjust "worse" ones, the default assumption seems to be that conflicting representations will exert equal and opposite influences upon each other.

It is possible, of course, that people are designed to achieve coherence among their mental states merely for the sake of coherence itself, without attempting to increase the average quality of the resulting representations. But as long as our minds are achieving coherence, shouldn't they hold relatively fixed those mental state representations that are most likely to be adaptive, while adjusting those that are most likely to be maladaptive? By analogy, consider a marketing company with one headquarters in New York and another in L.A. It would make the company more efficient and nimble to consolidate into a single headquarters. One possibility is to split the difference and relocate to Topeka – at least it would increase coherence! Another possibility, however, would be to carefully consider the performance of the New York and L.A. offices, and then to move the weaker headquarters to the location of the stronger one. This solution achieves coherence with a much more promising outlook for the bottom line. But it also makes greater demands on the decision maker. Like antecedent rationalization, it requires a method of prioritizing some representational exchanges over others.

R4.3. Revision: Coherence with priority

These commentaries offer two important amendments to the original theory of representational exchange: One mechanistic, and one functional. The mechanistic amendment is simply to point out that rationalization is not always ex post, but may often be ex ante, or a dynamic process in which coherence is achieved among many sorts of representations at any particular time. The functional amendment addresses the question at the heart of the target article: Why bother with rationalization? The target article offered two reasons; the commentaries suggest a third.

First, the target article suggested that we rationalize in order to improve reasoning, imparting truer beliefs and more useful (i.e., fit) desires. But this answer must be incomplete, because natural selection does not directly favor superior reasoning for its own sake; it only favors superior behavior. Why would we expect rationalization to improve the overall behavior of the organism (see **Berthelette & Kalbach**)? To the extent that the organism is already acting well based on some non-rational system, what use is there for improvement to its beliefs and desires?

The target article addressed this question in the broader context of representational exchange, offering a second kind of explanation. Different ways of representing information have different costs and benefits. Among the distinctive benefits of reasoning are its flexibility (i.e., the ability to rapidly adjust in light of new information) and its unique form of generalization (i.e., the ability to generate a policy in novel circumstances according to underlying

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

causal principles). Thus, extracting information from non-rational systems may enhance an organism's capacity for flexibility and generalization, improving its behavior overall. Meanwhile, other directions of representational exchange (e.g., from reasoning to habit, as during habitization) can leverage the distinct advantages of other forms of behavioral control (e.g., the speed and cognitive efficiency of habitual action).

The commentaries, however, draw our attention to another manner in which representational exchange can improve our behavior. If it can successfully prioritize from moment to moment which behavioral systems get to be the source versus the receiver of representational exchanges, this would ensure that representational exchange would lead not only to coherence among systems, but also improvement among them. This stands out as a key area of development for the theory of representational exchange. Existing theories of metacontrol provide a natural starting point.

R5. Social functions of rationalization

The theory of representational exchange proposes an "ultimate" or "computational"-level account of rationalization (Marr 1982; Tinbergen 1963). In other words, it is supposed to answer the question, "Why would natural selection favored rationalization?" Several commentaries also take up this question. Some extend the theory of representational exchange, while others propose alternatives to it. Nearly all of these contributions, however, focus on social functions of rationalization.

R5.1. Cohesion

Van Bavel, **Sternisko**, **Harris**, **& Robertson** (**Van Bavel et al.**), and **Gelpi, Cunningham, & Buchsbaum** (**Gelpi et al.**) propose that rationalization promotes social cohesion. When we rationalize the others' behavior, or collective cultural practices, we bring our beliefs and desires into closer alignment with that of others. The result is cohesion and a sense of shared identity, which, they propose, has direct personal benefits. As Gelpi et al. write, "Sharing a belief with those in one's community is therefore beneficial not only when (and because) that belief is true, but also when (and because) it provides an individual with the benefits of a group, such as a sense of belonging and easily accessible shared knowledge."

Certain versions of this thesis are closely related to the persuasion and blame-avoidance functions of rationalization discussed in the target article (**Altay & Mercier**, see also Haidt 2001; Mercier & Sperber 2011; Tedeschi et al. 1971; Tetlock 2002; Von Hippel & Trivers, 2011). For instance, suppose a political leader is caught performing an apparently corrupt action. In this case she and her supporters might rationalize her behavior by attempting to provide benign explanations for it (Altay & Mercier). But the commentaries also consider extensions of this thesis to many other kinds of cases: We might rationalize collective practices like holidays and sports; shared norms of how we spend time, organize families, and maintain friendships; and, perhaps above all, moral values (**Gelpi et al.** and **Graham**).

Van Bavel et al. argue that, when it supports social cohesion, "rationalization often facilitates distortion of rather than approximation to truth." **Altay & Mercier** and **Gelpi et al.** mostly share this view. In other words, theirs is an account of self-serving, pejorative rationalization (albeit one that serves an important social purpose). They do not consider it maladaptive, however, because it pays to get along with one's peers. (In contrast, **Graham** and **Laurin & Jettinghoff** raise the possibility that rationalizing collective action can sometimes be adaptive due to its self-benefitting nature. This possibility is considered at length in sect. R7.2.)

R5.2. Coordination

Saxe & Nettle, **Stanford et al.**, and **Levy** propose an alternative social function of rationalization: Facilitating coordinated joint action. Humans are nature's paramount cooperators; we have an unmatched ability to flexibly coordinate our behavior to accomplish joint goals. We succeed in part because we can predict each other's behavior. We mostly predict each other's behavior by assuming rational planning and choice (Baker et al. 2009; Dennett 1987). Thus, the commentators propose, it will be easier to explain and predict another person's action if it exhibits consistent adherence to norms of rationality across time. Rationalization may be useful because it imposes such consistency. As Levy writes, "By making ourselves predictable, we enable more efficient cooperation, which is essential for the flourishing of social animals like us (Tomasello 2014). If each of us can predict how the others will behave, we can more efficiently play our part in joint actions, without interfering with one another or introducing redundancies."

If anything, this understates the case. Successful joint action often requires not only that we model each other's intentions, but also that we adopt joint intentions, in which the behavior of each partner is generated by shared goals and common knowledge (Kleiman-Weiner et al. 2016; Tomasello 2005). Possibly, then, rationalization allows us to translate the behavioral prescriptions of non-rational systems into a format amenable to joint intentional action.

R5.3. Communication

Levy and **Pärnamets et al.** consider a final social function of rationalization: to foster communication between people. Whether or not reasoning is the language of thought, it is certainly the language of language. We explain ourselves with words, and these words typically express our beliefs, desires, plans, and reasons. Perhaps this is because beliefs, desires, plans, and reasons are especially effective kinds of representations for communication.

It is obvious, for instance, that one person cannot describe their instinct (i.e., innate behavior) to another person and thereby cause the other person to acquire an *instinct*. Rather, some change of representational format is required. In theory, a person could express their instinct (or habit, or norm) simply as a behavioral prescription. But it will often be much more useful to learn the reasons behind a person's policy than to simply learn the policy itself and be forced to copy it. This is because each of us occupies different circumstances, and so the optimal policy for one person may be suboptimal for another. Suppose, for instance, you love peanuts, but I am allergic to them. If you tell me, "Eat these cookies!" I may be substantially worse off than if you tell me, "I eat these cookies because I love the peanuts in them." Rationalizing prior to communication presents information in a format such that a person can flexibly adapt socially learned information to their own unique circumstances.

The target article contrasted two possible pathways of social learning (see Fig. 3 of the target article). In the first, a person acts, an observer engages in theory of mind to extract the beliefs and desires implicit in the action, and the observer finally adopts those beliefs and desires herself (i.e., by informational conformity). In the second, a person acts, an observer directly copies

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

the behavior (i.e., by normative conformity) and, upon rationalizing her own behavior, the observer finally extracts its implicit beliefs and desires and adopts them. The key insight of **Levy** and **Pärnamets et al.** is to add a third pathway: A person acts, he rationalizes his own behavior (extracting beliefs and desires from it), and he verbally reports these to an observer who then adopts them. This is a promising addition to the framework.

R5.4. Revision: The social dimension of representational exchange

Each of these reviews posits that rationalization has important and diverse social benefits, fostering cohesion, coordination, and communication. Some of these functions, such as group cohesion, share little overlap with the theory of representational exchange. They more likely explain self-serving rationalization. Others of these, such as coordination and communication, are likely explanations of self-benefitting rationalization. They show that effective representational exchange improves not just our personal thoughts and actions, but also our collective ones.

What is most remarkable about this set of commentaries, however, is their implied commitment to a bold thesis: That social benefits are not just an important explanation of rationalization, but also of reason itself. In fact, there is an even more extreme version of this hypothesis, according to which reason (beliefs, desires, plans, etc.) plays no important role in guiding our behavior, but is instead a socially useful way of talking about our behavior. The next section considers several proposals that engage this possibility.

R6. Do we reason at all?

The target article argued that often the function of rationalization is to improve reasoning by improving its raw materials: The beliefs and desires that it assembles into plans. **Levy**, **Roskies**, and **Veit, Dewhurst, Dolega, Jones, Stanley, Frankish, & Dennett** (**Veit et al.**) question whether humans actually reason – in other words, whether we hold beliefs and desires and use them to make plans. None of these authors dispute that we talk in terms of reasons, beliefs, desires, and plans. And, they agree that such talk has meaning and utility. But, they suggest, maybe it's just talk. (Some endorse this possibility more strongly than others). The target article described rationalization as a "useful fiction," but is the real fiction reason itself?

According to these critiques, people exhibit "flexible response[s] to environmental and internal information," but they do so in ways that do not depend on "explicit representations" of belief and desire (**Levy**). Rather, what exist are "some sort of proto-mental states" (**Veit et al.**), the "great majority of these states are dispositional and may never be explicitly represented" (Levy). **Roskies** offers a clear introduction to these ideas, employing the helpful example of a thermostat: "To the degree that we can predict and explain a system's behavior by imputing folk psychological states [e.g., beliefs, desires, plans, and reasons] to that system, so we are warranted in that imputation. For Dennett [1987], even simple systems such as thermostats are legitimate targets of the intentional stance, even though no standard realist would support the view that thermostats have mental states." The proposed function of the intentional stance, not dissimilar to Bem's (1972), is not to improve one's own reasoning, but rather to allow us to describe, predict, and explain our own and others' actions.

There is much to like in these commentaries, but also something amiss. Surely, it is often useful to impute reasons to psychological systems that do not reason. This is the essence of the target article's model of rationalization as representational exchange. And, as discussed in section R4, rationalization has benefits not only for the individual who rationalizes, but also for those who depend upon her for collaboration and communication. But here is the key question: Do these facts explain away reasoning itself? If we accept that our experience of having reasoned in sometimes a fiction, do we have any reason to accept that it is ever a reality?

We do. As defined in the target article, to act by reasoning is to estimate the expected value of candidate actions by consulting a generative model of their consequences, and to then choose actions proportional to their expected value (following, e.g., Daw & Dayan 2014). There is ample evidence that people learn and represent generative causal models (Buckner et al. 2008; Gershman 2018; Gläscher et al. 2010) and can use these to assess the likely consequences of candidate actions (Deserno et al. 2015; Doll et al. 2015; Simon & Daw 2011; Vikbladh et al. 2019). By considering these consequences in light of our goals our preferences, we can estimate the instrumental value of candidate actions (Bakkour et al. 2019; Daw et al. 2011; Jones et al. 2012). These value estimates reliably influence choice (Cushman & Morris 2015; Daw et al. 2011; Deserno et al. 2015; Doll et al. 2015; Simon & Daw 2011). These are precisely the kinds of representations and computations that can be improved by representational exchange.

In summary, these commentaries emphasize the striking and illuminating fact that we can profitably understand a thermostat by treating it as rational, even if it is not (i.e., by adopting the intentional stance). This immediately suggests a subversive and alluring possibility: Maybe, despite feeling and talking like rational creatures, we are more like very complicated thermostats. Yet, as an empirical matter, we are not. When a thermostat adjusts the temperature, it is not by encoding a generative model of the world, computing the likely consequences of various actions, and then choosing actions proportional to the resulting instrumental value estimates. But, when a human adjusts the thermostat, it often is.

R7. Making meaning

We cannot always know the causes of our behavior, but rather must infer them (Bem 1972; Nisbett & Wilson 1977). We do this in part because it brings order and meaning in our lives. Many theories posit that we search for order and meaning because it makes us feel better, which is a proximate explanation (Baumeister 1991; Hasselkus 2011; Reker et al. 1987; Webster & Kruglanski 1994). The theory of representational exchange provides a complementary ultimate explanation: The meaning we make out of our actions, attitudes, and emotions (**Railton**) renders useful information available for reasoning. A final set of commentaries takes this idea of meaning-making as its point of departure. The authors observe how phenomena as diverse as daydreaming ("everyday thought experiments": **Ullman**), system justification (**Laurin & Jettinghoff**), and ideological commitment (**Graham**) can be understood as attempts to find meaning, breaking new ground for the theory of representational exchange. They make a convincing case that these are powerful and constant forces in our thinking. They also offer a cautionary note, questioning what happens when even self-benefitting rationalization is misapplied (**Weinberg & Weinberg, Pärnamets et al.**, Graham).

R7.1. Rationalization and personal meaning

People daydream about all kinds of fantastical things. They love it. They'll do it on their own, and as **Ullman** shows, they'll happy to do it if you ask. This is remarkable because many of the questions posed by McCoy et al. (2019, p. 248) are patently outlandish:

> Imagine that aliens come down to Earth, and give you the option to go with them on their travels throughout the universe. The aliens are friendly and honest, and tell you that you would see amazing things on your travels with them if you decide to go with them.
>
> If you decide to go, you will have a week to say goodbye to your family and friends. Once you leave, you will never again return to Earth, nor be able to communicate with people on Earth.
>
> Do you go?

Different people give different answers, and people are bad at guessing what others will say, yet most people report being highly confident in their own personal answers. Not only do people say they enjoy thinking about things like this, they say that they *learned* something from it – mostly things about themselves:

> That I am more risk averse when faced with transformative decisions.
>
> That I am attached to my actual, earthly existence more than I might have thought otherwise.
>
> That I value personal relationships more than I thought.

These findings comport with a large literature showing that when people have downtime, their minds turn toward thoughts about imagined possibilities for themselves, in the past and in the future (Buckner et al. 2008).

As **Ullman** notes, the theory of representational exchange "can help explain why such everyday thought experiments are informative, and also why people like to engage in them." When people surprise themselves with their own confident answers to hypothetical questions, it means that they are discovering implicit information in their automatic reactions, including emotions (**Railton**) – information that often contradicts their explicitly held beliefs and values. They consider this information valuable and *true* – they take it to reflect important insights about who they are and how they ought to feel. By organizing our impulses around meaning, purpose, and narrative, we unpack implicit information from non-rational systems and make it available to improve subsequent reasoning.

R7.2. Rationalization and social meaning

Although the target article focused mostly on the rationalization of one's own actions, **Graham** and **Laurin & Jettinghoff** propose a key extension, and Graham writes: "Processes like rationalization often occur at the collective levels of groups, societies, and cultures. Collective rationalization – whereby a group's collective action leads them to update or create new shared beliefs and desires – may be rational as well. Just as individual rationalization can extract useful information from non-rational sources like instincts and habits, communal processes can cohere moral intuitions and norms into the shared 'useful fictions' of shared moral narratives." Graham's commentary focuses on ideology: A shared sense of meaning – an interconnected set of beliefs and values that give a sense of coherence, explanation and purpose to groups (Jost 2006). Laurin & Jettinghoff's commentary focuses on system justification: Our tendency to assume that prevailing social structures are good. They agree that we seek not only to ascribe meaning to our individual actions, but also to collective cultural practices.

Laurin & Jettinghoff wonder, however, why we would be justified in not just inferring but actually adopting the beliefs and desires that we attribute to others. In other words, why drink the collective Kool-Aid? "If people were merely trying to identify beliefs and desires" that make sense of other's actions, "there would be no reason to expect system justifying rationales to predominate" over system-undermining ones. Why assume the system is good?

Different commentaries suggest different potential answers to this problem. **Gelpi et al.** appeal to cognitive efficiency: "Trusting the beliefs shared by one's social group – and 'outsourcing' one's own cognition to depend on knowledge held by others in their community – can also reduce the need to engage in cognitively effortful reasoning … (Sloman & Rabb 2016)." In contrast, **Laurin & Jettinghoff** consider the possibility that we do not rationalize the system itself, but rather our own apparent acceptance of it. Perhaps, then, system justification is really justification of one's own complacency. Both of these explanations probably capture a good part of the truth, along with the possibility that the rationalization of collective behavior fosters social cohesion (**Van Bavel et al.** and Gelpi et al.).

But there is also a far simpler answer: We rationalize collective cultural practices, adopting their implied beliefs and values, because collective cultural practices are usually pretty good. Just as genes are subject to genetic evolution, collective practices are subject to cultural evolution (Boyd et al. 2011). Thus, if you observe that most people are doing something (or that the cultural system is organized in some way), that thing is more likely to be adaptive than maladaptive. This does not imply that the current norm is the best that could possibly exist, any more than we should suppose that a sparrow's wings are the best wings that a sparrow could possibly have. Rather, it implies that adhering to the norm is probably better than ignoring it, just as the sparrow is better off with her wings than without them.

Confronted with a claim like this – *if most people do it, it's probably adaptive* – our minds naturally leap to counterexamples. Wasn't it for good reason that our mothers scolded, "If everybody jumped off a bridge, would you?" After all, people do all sorts of terrible things. Most people eat too much, don't get enough exercise, are cognitively and socially biased, and so forth. There was a time and a place when most people thought the Earth was flat; when most people smoked cigarettes; and when most people thought it was acceptable to whip your child, your wife, and your slave. Human progress has always relied on those who say, "I can do better than what most people are doing," and it always will.

These counterexamples are important and inescapable. But now imagine what it would mean to abandon our cultural inheritance entirely. Imagine a child who grows up without any inclination to think or act like others. Given soup, she would have no inclination to use a spoon just because her big sister does; rather, she would *reason* about whether a spoon works for her. The same goes for forks, manners, values, laws, school, sports, holidays, fashions, jobs, and so on. Shown a recipe for bread, she would not assume there was anything useful about the yeast, the kneading, the rise, or even the oven. She would just rely on her own knowledge and sound reasoning to deduce the proper manner of baking. Now, whose loaf would you rather eat: Hers or an obedient apprentice's?

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Our collective cultural practices are a recipe for life. It is not a perfect recipe – in fact, you can almost certainly improve it with some tweaks. But it is unlikely that you will do better by ignoring the recipe altogether. More likely, any improvements you make will necessarily build upon the cumulative wisdom of generations of cooks that have come before you. (In Newton's terms, seeing further than giants requires not just extending your legs, but also climbing on their shoulders). Sometimes most people do the wrong thing, but these counterexamples miss the forest for the trees – nearly everything we do right was learned, in part, from those around us (Boyd et al. 2011). This accords with a basic impulse of conservatism in political theory and philosophy (e.g., Hayek 1973): Tradition may require revision or rejection, but it at least deserves respect.

Or, at least, it does if you are interested in fitness within the framework of biological and cultural evolution. As scientists attempting to understand the structure of the human mind – in particular, our propensity to rationalize both personal and collective actions – this is the kind of fitness we are interested in as a descriptive theory. But what was fit for the past may be unfit for the present, and what is fit for nature may be unfit for ethics.

R7.3. Spoiled Kool-Aid: Is rationalization outdated?

While accepting that rationalization may have an adaptive rationale, **Weinberg & Weinberg, Pärnamets et al.**, and **Graham** ask: Is it the right thing to do? They offer three reasons to think it is not.

The first is familiar: At least some rationalization is self-serving, and it will generally lead us away from the truth. While it may have some adaptive benefits when it works properly (such as cohesion, as noted by **Van Bavel et al.**), **Graham** documents the pernicious effects when it goes awry, such as the justification of prejudice or even violence toward outgroups.

Graham goes further, however, by questioning even self-benefitting rationalization. He accepts that by adopting the values implied by instincts and norms, we are adopting values shaped by adaptive processes – biological and cultural evolution, respectively. These may be fit, he notes, but are they morally right? This is an important question; indeed, it is among the most important questions we face. If rationalization captures the conservative impulse to respect traditions, we must confront the fact that many human traditions are morally abhorrent. So, too, is blind respect for them.

Weinberg & Weinberg and **Pärnamets et al.** offer a final reason to distrust self-benefitting rationalization: While it may have been adaptive when it evolved, perhaps it is poorly suited to present circumstances. We evolved at a time when the world was relatively more static. Thus, norms, instincts, and habits shaped by the past were usually adaptive in the present. And, in turn, implicit beliefs and desires extracted from these sources were generally useful. Yet, today the world changes at a faster pace environmentally, technologically, and socially. This drives our instincts, norms, and habits out of date more quickly than ever before. (And while the commentaries do not mention it, it is also likely that our beliefs are truer, and capacity for reason keener, than ever before.) If natural selection struck a balance between non-rational systems and reasoning when rationalization evolved, perhaps it is the wrong balancing point today. This point is worth careful consideration. If indeed the balancing point of representational exchange is innate and shaped by biological natural selection, then it is hard to imagine that it is optimal today. Yet, as discussed in section R4, the balancing act between reason and intuition may be carried out online by metacontrol processes that are relatively more sensitive to present circumstances.

R7.4. Revision: The scope and challenge of representational exchange

This final group of commentaries offers two important revisions to the target article. First, they broaden the reach of the representational exchange to personal and collective self-understanding: The attempt to find meaning in our lives. The target article talked a lot about extracting information, but not at all about making meaning. Yet, the search for narrative, structure, and meaning in our lives is a perfect example of representational exchange. It may be a fiction to suppose that our individual or collective behavior already *has* meaning, as if organized by an intelligent agent with a specific plan in mind. But if it is a fiction, it is surely a useful one. We stand to learn a tremendous amount by assuming that the way each and all of us do things is, while not perfect, probably very good. Second, however, they sound a cautionary note even for self-benefitting rationalization. It may be adaptive, but is it good? They provide ample reasons for skepticism.

On their face, these commentaries seem to strike two discordant notes. On the one hand, representational exchange is at the heart of our search for meaning. On the other, it may often be a misguided search. Yet, in fact, these notes combine to play a familiar theme: the familiar struggle to balance reasoning against intuition – head against heart.

R8. The human problem

The Enlightenment bestowed a durable understanding of the human problem: To act rationally and, therefore, by reasoning. Summarizing this neoclassical spirit, Pinker (2018, p. 8) writes: "If there's anything the Enlightenment thinkers had in common, it was an insistence that we energetically apply the standard of reason to understanding our world, and not fall back on generators of delusion like faith, dogma, revelation, authority, charisma, mysticism, divination, visions, gut feelings or the hermeneutic parsing of sacred texts."

Although not our only understanding of the human problem, this is an influential one. It aligns with a strong tradition of psychological research that reifies reasoning, organizing it around two normative standards. First, one ought to establish beliefs by the rational application of Bayesian principles. Second, one ought to choose actions by expected value maximization. Against the background of these standards, the behavioral prescriptions of norms, instincts, and habits are defined as biases, mostly to be understood in light of their systematic deviations from reasoning. If they can be justified at all, it is either on the grounds that they are more computationally efficient (Griffiths et al. 2015) or have indirect social benefits (Mercier & Sperber 2011; Tedeschi et al. 1971; Von Hippel & Trivers 2011). These ancillary benefits may compensate for what are otherwise strictly worse outcomes: mistaken beliefs, misguided desires, and maladaptive choices.

Thus, according to the strictest Enlightenment ideals, we must purge thought and action of the influence of non-rational systems. Pinker (2018, pp. 8–9), again: "Thinkers such as Kant, Spinoza, Hobbs, Hume and Smith were inquisitive psychologists and all

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

too aware of our irrational passions and foibles. They insisted that it was only by calling out the common sources of folly that we could hope to overcome them. The deliberate application of reason was necessary precisely because our common habits of thought are not particularly reasonable." This is why we might characterize reasoning not as the human condition, but rather the human problem. It is a problem because, while reasoning is presumed to be superior, it also hard to do. If our highest human calling is self-improvement, then we must apply the force of will to achieve the purity of reasoned thinking.

In its purest form, this Enlightenment vision possesses the allure of simplicity, the veneer of logic, and the tragic myopia of self-certainty. Human reasoning is, after all, exquisitely fallible. Our beliefs are often incorrect and always incomplete. Our desires may internally inconsistent, interpersonally irreconcilable, and occasionally detrimental to our well-being. And our capacity for rational planning is severely constrained by time and cognitive bandwidth. None of this implies that reason is bad; to the contrary, it is indispensable. But it is fallible, and sometimes it can be improved by attending to non-rational adaptive mechanisms. Whether through the genetic inheritance of our instincts, the cultural inheritance of our norms, or the experiential inheritance of our habits, we are endowed with guides to action shaped by powerful forces of trial-and-error adaptation.

Thus, to act truly rationally cannot be to act only by reasoning. Rather, even if we accept the Enlightenment commitment to rational action – *especially* if we accept it – then the basic human problem is not to isolate reason from non-rational processes, but to improve it by considering them. This, of course, is Jason Bourne's problem: To glean from his programmed abilities whatever information serves his present interests. In this effort, rationalization is an essential tool. It is not a matter base self-enhancement or even one of elevated self-understanding. It is, at heart, a project of self-improvement – surely the highest human calling of all.

Note

[1] In the target article I wrote of this finding: "These cases are hard to see as anything but gross errors." Some of the commentaries interpreted this to mean a *large* error, but it was intended instead in the sense an *undeniable* error. (See, e.g., the first entry of the Merriam Webster Online definition of "gross": "glaringly noticeable usually because of inexcusable badness or objectionableness // 'a gross error.'" https://www.merriam-webster.com/dictionary/gross) I also wrote that, "Sensibly or not, people rationalize all the time." Some of the commentaries interpreted this rather literally, almost as if I had meant, "At most points in time, people are rationalizing." Rather, I intended to convey, "People rationalize a lot more than you might have thought, given that it is typically considered 'not sensible.'" By analogy, "The president plays golf all the time" is meant to convey that she plays a lot more golf than one might have supposed, given the demands of her office.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Abelson, R. P. (1983) Whatever became of consistency theory? *Personality and Social Psychology Bulletin* **9**:37–64. [DS]

Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. (1985) A learning algorithm for Boltzmann machines. *Cognitive Science* **9**(1):147–69. [aFC]

Adler, J. E. & Rips, L. J. (eds.). (2008) *Reasoning: Studies of human inference and its foundations*. Cambridge University Press. [AD]

Ajzen, I. (1985) From intentions to actions: A theory of planned behavior. In: *Action control: From cognition to behavior*, ed. J. Kuhl & J. Beckman, pp. 11–39. Springer-Verlag. [WT]

Ajzen, I. & Fishbein, M. (2005) The influence of attitudes on behavior. In *The handbook of attitudes* (pp. 173–221). Erlbaum. [AD]

Akerlof, G. & Yellen, J. (1985) A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* **100**(5):823–38. [SEW]

Alicke, M. D. (2000) Culpable control and the psychology of blame. *Psychological Bulletin* **126**(4):556–74. [aFC]

Alloy, L. B. & Abramson, L. Y. (1979) Judgments of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General* **108**(4):441–85. [JQ-D]

American Psychiatric Association. (2000) *Diagnostic and statistical manual of mental disorders, 4th ed., text rev. (DSM-IV-TR)*. American Psychiatric Association. [SBr]

Andrews, G., Singh, M. & Bond, M. (1993) The defense style questionnaire. *Journal of Nervous and Mental Disease* **181**(4):246–56. [SBr]

Armitage, C. J. & Conner, M. (2001) Efficacy of the theory of planned behavior: A meta-analytic review. *British Journal of Social Psychology* **40**:471–99. [WT]

Arnold, M. B. (1960) *Emotion and personality*. Columbia University Press. [aFC]

Aronson, E. (1968) Dissonance theory: Progress and problems. In: *Theories of cognitive consistency: A sourcebook*, ed. R. P. Abelson, E. E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg & P. H. Tannenbaum, pp. 5–27. Rand-McNally. [aFC]

Aronson, E. (1969) The theory of cognitive dissonance: A current perspective. *Advances in Experimental Social Psychology* **4**:1–34. [JQ-D]

Aronson, E. (1992) The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry* **3**(4):303–11. [JQ-D]

Aronson, E. & Mills, J. (1959) The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology* **59**:177–81. [JQ-D]

Auer, P., Cesa-Bianchi, N. & Fischer, P. (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47**(2–3):235–56. [TDU]

Badre, D. & Nee, D. E. (2017) Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences* **22**(2):170–88. [aFC]

Bago, B. & De Neys, W. (2019) The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, **3**, 257–99. [WDN]

Bahaddin, B., Weinberg, S., Luna-Reyes, L. & Andersen, D. (2019) Building a bridge to behavioral economics: Countervailing cognitive biases in lifetime saving decisions. *System Dynamics Review* **35**(3):187–207. [SEW]

Bahamondes, J., Sibley, C. G. & Osborne, D. (2019) "We look (and feel) better through system-justifying lenses": System-justifying beliefs attenuate the well-being gap between the advantaged and disadvantaged by reducing perceptions of discrimination. *Personality and Social Psychology Bulletin* **45**(9):1391–1408. [KL]

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F. & Volfovsky, A. (2018) Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* **115**:9216–21. [JJVB]

Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. (2017) Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* **1**(4):1–10. https://www.nature.com/articles/s41562-017-0064 [TDU]

Baker, C. L., Saxe, R. & Tenenbaum, J. B. (2009) Action understanding as inverse planning. *Cognition* **113**(3):329–49. [arFC, TDU]

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H., Reid, A., Verfaellie, M., Shadlen, M. N., & Shohamy, D. (2019) The hippocampus supports deliberation during value-based decisions. *eLife* **8**. doi:10.7554/elife.46080. [rFC]

Baron-Cohen, S. (1995) *Mindblindness: An essay on autism and theory of mind*. MIT Press. [aFC]

Batchelder, W. H. & Alexander, G. E. (2012) Insight problem solving: A critical examination of the possibility of formal theory. *Journal of Problem Solving* **5**(1):56–100. [aFC]

Baumard, N., André, J.-B. & Sperber, D. (2013) A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences* **36**(2):59–122. [PKS]

Baumeister, R. F. (1982) A self-presentational view of social phenomena. *Psychological Bulletin* **91**(1):3–26. [SA]

Baumeister, R. F. (1991) *Meanings of life*. Guilford Press. [rFC]

Baumeister, R. F. & Cairns, K. J. (1992) Repression and self-presentation: When audiences interfere with self-deceptive strategies. *Journal of Personality and Social Psychology* **62**(5): 851. [SA]

Bayer, H. M. & Glimcher, P. W. (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**(1):129–41. [aFC]

Beck, A. T. (2008) The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry* **165**(8):969–77. [JQ-D]

Beggan, J. K. (1992) On the social nature of nonsocial perception: The mere ownership effect. *Journal of Personality and Social Psychology* **62**(2):229–37. [aFC]

Bellman, R. (1954) The theory of dynamic programming. *Bulletin of the American Mathematical Society* **60**(6):503–15. [aFC]

Bem, D. J. (1967) Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review* **74**(3):183–200. [aFC]

Bem, D. J. (1972) Self-perception theory. In: *Advances in experimental social psychology, vol. 6*, pp. 1–62. Elsevier. [arFC, WT]

Benartzi, S. & Thaler, R. (1995) Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics* **110**(1):73–92. [SEW]

Benjamin, D. (2018) Errors in probabilistic reasoning and judgment biases. National Bureau of Economic Research Working Paper No. 25200. https://doi.org/10.1016/bs.hesbe.2018.11.002. [SEW]

Bentler, P. M. & Speckart, G. (1981) Attitudes "cause" behaviors: A structural equation analysis. *Journal of Personality and Social Psychology* **40**:226–38. [WT]

Berkowitz, L. & Devine, P. G. (1989) Research traditions, analysis, and synthesis in social psychological theories: The case of dissonance theory. *Personality and Social Psychology Bulletin* **15**:493–507. [DS]

Bicchieri, C. (2005) *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press. [PP]

Blais, M. A., Conboy, C. A., Wilcox, N. & Norman, D. K. (1996) An empirical study of the DSM-IV Defensive Functioning Scale in personality disordered patients. *Comprehensive Psychiatry* **37**(6):435–40. doi: 10.1016/S0010-440X(96)90027-9. [SBr]

Bortolotti, L. (2015) The epistemic innocence of motivated delusions. *Consciousness and Cognition* **33**:490–99. [JE]

Botvinick, M. & Weinstein, A. (2014) Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B* **369** (1655):20130480. http://dx.doi.org/10.1098/rstb.2013.0480 [aFC]

Botvinick, M. M. (2008) Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences* **12**(5):201–208. [aFC]

Boyd, R., Richerson, P. J. & Henrich, J. (2011) The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences USA* **108**:10918–25. [arFC]

Brehm, J. W. (1956) Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology* **52**(3):384–89. [aFC, DS, JQ-D, TS]

Brescoll, V. L., Uhlmann, E. L. & Newman, G. N. (2013) The effects of system-justifying motives on endorsement of essentialist explanations for gender differences. *Journal of Personality and Social Psychology* **105**:891–908. [WT]

Briley, D. A., Morris, M. W. & Simonson, I. (2000) Reasons as carriers of culture: Dynamic versus dispositional models of cultural influence on decision making. *Journal of Consumer Research* **27**(2):157–78. [SA]

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence* **47**(1–3):139–59. [PP]

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S. & Colton, S. (2012) A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games* **4**(1):1–43. [aFC]

Brownhalls, J., Duffy, A., Eriksson, L. & Barlow, F. K. (2019) Reintroducing rationalization: A study of relational goal pursuit theory of intimate partner obsessive relational intrusion. *Journal of Interpersonal Violence*. (Advance online publication) doi: 10.1177/0886260518822339. [SBr]

Brownstein, A. L. (2003) Biased predecision processing. *Psychological Bulletin* **129**:545–68. [DS]

Bruner, J. (1957) Discussion. In: *Contemporary approaches to cognition: A symposium held at the University of Colorado*, ed. J. S. Bruner, pp. 151–56. Harvard University Press. [DS]

Bryant, E., Schimke, E. B., Nyseth Brehm, H. & Uggen, C. (2018) Techniques of neutralization and identity work among accused genocide perpetrators. *Social Problems* **65**(4):584–602. doi: 10.1093/socpro/spx026. [SBr]

Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. (2008) The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*. **1124**:1–38. [rFC]

Buckner, R. L. & Carroll, D. C. (2007) Self-projection and the brain. *Trends in Cognitive Sciences* **11**(2):49–57. [aFC]

Calvete, E. (2008) Justification of violence and grandiosity schemas as predictors of antisocial behavior in adolescents. *Journal of Abnormal Child Psychology* **36**(7):1083–95. doi: 10.1007/s10802-008-9229-5. [SBr]

Carlson, K. A. & Russo, J. E. (2001) Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied* **7**(2):91–103. [DS]

Carpenter, S. M., Yates, J. F., Preston, S. D. & Chen, L. (2016) Regulating emotions during difficult multiattribute decision making: The role of pre-decisional coherence shifting. *PLOS ONE* **11**:e0150873. [DS]

Carruthers, P. (2013) *The opacity of mind: An integrative theory of self-knowledge*. Oxford University Press. [NL]

Caruso, E. M., Shapira, O. & Landy, J. F. (2017) Show me the money: A systematic exploration of manipulations, moderators, and mechanisms of priming effects. *Psychological Science* **28**:1148–59. [WT]

Chaxel, A. S., Russo, J. E. & Kerimi, N. (2013) Preference-driven biases in decision makers' information search and evaluation. *Judgment and Decision Making* **8**(5):561–76. [DS]

Christensen-Szalanski, J. J. & Willham, C. F. (1991) The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes* **48**(1):147–68. [aFC]

Churchland, P. M. (1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* **78**(2):67–90. [ALR, WV]

Churchland, P. S. (1986) *Neurophilosophy: Toward a unified science of the mind/brain*. MIT Press. [WV]

Cialdini, R. B. & Goldstein, N. J. (2004) Social influence: Compliance and conformity. *Annual Review of Psychology* **55**:591–621. [aFC]

Cialdini, R. B. & Trost, M. R. (1998) Social influence: Social norms, conformity and compliance. In: *The handbook of social psychology*, ed. D. T. Gilbert, S. T. Fiske & G. Lindzey, pp. 151–92. McGraw-Hill. [aFC]

Clement, J. J. (2009) The role of imagistic simulation in scientific thought experiments. *Topics in Cognitive Science* **1**(4):686–710. [TDU]

Cohen, J. (1992) A power primer. *Psychological bulletin* **112**(1):155–59. [TS]

Colvin, C. R., Block, J. & Funder, D.C. (1995) Overly-positive self evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology* **68**:1152–62. [WT]

Cooper, J. (2007) *Cognitive dissonance: 50 years of a classic theory*. SAGE. [NL]

Cromwell, P. & Thurman, Q. (2003) the devil made me do it: use of neutralizations by shoplifters. *Deviant Behavior* **24**:535–50. doi: 10.1080/01639620390225859. [SBr]

Cushman, F. (2013) Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review* **17**(3):273–92. [PP]

Cushman, F. & Paul, L. (in press) Are desires interdependent? In: *The Oxford handbook of moral psychology*, ed. J. Doris & M. Vargas. Oxford University Press. [aFC]

Cushman, F. A. & Morris, A. (2015) Habitual control of goal selection in humans. *Proceedings of the National Academy of Science USA* **112**(45):13817–22. [arFC]

Dahl, A. (2017) Ecological commitments: Why developmental science needs naturalistic methods. *Child Development Perspectives* **11**(2):79–84. Available at: https://doi.org/10.1111/cdep.12217. [AD]

Dahl, A. (2019) The science of early moral development: On defining, constructing, and studying morality from birth. *Advances in Child Development and Behavior* **56**:1–35. Available at: https://doi.org/10.1016/bs.acdb.2018.11.001. [AD]

Dahl, A., Gingo, M., Uttich, K. & Turiel, E. (2018) Moral reasoning about human welfare in adolescents and adults: Judging conflicts involving sacrificing and saving lives. *Monographs of the Society for Research in Child Development, Serial No. 330*, **83**(3):1–109. [AD]

Dahl, A. & Killen, M. (2018) Moral reasoning: Theory and research in developmental science. In: *The Stevens' handbook of experimental psychology and cognitive neuroscience*, 4th edition, ed. J. T. Wixted & S. Ghetti, pp. 323–53. Wiley. [AD]

Dahl, A. & Waltzer, T. (2018) Moral disengagement as a psychological construct. *American Journal of Psychology* **131**:240–46. [AD]

Daniel, K. & Hirshleifer, D. (2015) Overconfident investors, predictable returns, and excessive trading. *Journal of Economic Perspectives* **29**(4):61–88. [SEW]

Daniels, N. (2003) Reflective equilibrium. In: *Stanford encyclopedia of philosophy* (fall 2018 edition), ed. Edward N. Zalta. Available at: https://plato.stanford.edu/entries/reflective-equilibrium/#pagetopright. (Online publication) [aFC, DS]

Davidson, D. (1970) How is weakness of will possible? In: *Moral concepts*, ed. J. Feinberg, pp. 93–113. Oxford University Press. [SBe]

Davidson, T. J., Kloosterman, F. & Wilson, M. A. (2009) Hippocampal replay of extended experience. *Neuron* **63**(4):497–507. [aFC]

Daw, N. D. & Dayan, P. (2014) The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369** (1655):20130478. [arFC]

Daw, N. D. & Doya, K. (2006) The computational neurobiology of learning and reward. *Current Opinion in Neurobiology* **16**(2):199–204. [aFC]

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**(6):1204–15. [arFC]

Daw, N. D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* **8**(12):1704. [arFC]

Dawson, L. L. (1999) When prophecy fails and faith persists: A theoretical overview. *Nova Religio: The Journal of Alternative and Emergent Religions* **3**(1):60–82. Available at: https://doi.org/10.1525/nr.1999.3.1.60. [NL]

Dayan, P. (2012) How to set the switches on this thing. *Current Opinion in Neurobiology* **22**(6):1068–74. [aFC]

Dayan, P. & Berridge, K. C. (2014) Model-based and model-free Pavlovian learning: Revaluation, revision, and revelation. *Cognitive and Affective Behavioral Neuroscience* **14**:473–92. [PR]

D'Cruz, J. (2015) Rationalization, evidence, and pretense. *Ratio* **28**(3):318–31. Available at: https://doi.org/10.1111/rati.12072. [JD]

DeKay, M. L. (2015) Predecisional information distortion and the self-fulfilling prophecy of early preferences in choice. *Current Directions in Psychological Science* **24**(5):405–11. [DS]

DeKay, M. L., Patiño-Echeverri, D. & Fischbeck, P. S. (2009) Distortion of probability and outcome information in risky decisions. *Organizational Behavior and Human Decision Processes* **109**(1):79–92 [DS]

De Neys, W. (ed.). (2017) *Dual process theory 2.0*. Routledge. [WDN]

Dennett, D. C. (1987) *The intentional stance*. MIT Press. [ALR, arFC, WV]

Dennett, D. C. (1991a) *Consciousness explained*. Penguin. [PP]

Dennett, D. C. (1991b) Real patterns. *The Journal of Philosophy* **88**(1):27–51. [PP]

Dennett, D. C. (1991c) Two contrasts: Folk craft versus folk science, and belief versus opinion. In: *The future of folk psychology: Intentionality and cognitive science*, ed. J. D. Greenwood, pp. 135–48. Cambridge University Press. [PP]

Dennett, D. C. (1996) *Kinds of minds: Toward an understanding of consciousness*. Basic Books. [PP]

Dennett, D. C. (2011) Homunculi rule: Reflections on Darwinian populations and natural selection by Peter Godfrey-Smith. *Biology & Philosophy* **26**:475–88. [WV]

Dennett, D. C. (2017) *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company. [WV]

Dennett, D. C. (2019) Clever evolution [Review of Samir Okasha's *Agents and goals in evolution*, Oxford University Press, 2018]. *Metascience* **28**(3):355–58. (Advance online publication). Available at: https://doi.org/10.1007/s11016-019-00450-w. [WV]

Deserno, L., Huys, Q. J. M., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., Dolan, R. J., Heinz, A. & Schlagenhauf, F. (2015) Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences USA* **112**(5):1595–1600. [rFC]

Deutsch, M. & Gerard, H. B. (1955) A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology* **51**(3):629–36. [aFC]

Devine, P. G. (1989) Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* **56**(1):5–18. [aFC]

deYoung, m. (1989) The world According to NAMBLA: Accounting for deviance. *Journal of Sociology & Social Welfare* **16**:111–26. [SBr]

Dezfouli, A. & Balleine, B. W. (2012) Habits, action sequences and reinforcement learning. *European Journal of Neuroscience* **35**(7):1036–51. [arFC]

Dezfouli, A. & Balleine, B. W. (2013) Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLOS Computational Biology* **9**(12):e1003364. [arFC]

Dezfouli, A., Lingawi, N. W. & Balleine, B. W. (2014) Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**(1655):20130482. [arFC]

Diehl, M., Chui, H., Hay, E. L., Lumley, M. A., Gruhn, D. & Labouvie-Vief, G. (2014) Change in coping and defense mechanisms across adulthood: longitudinal findings in a European American sample. *Developmental Psychology* **50**(2): 634–48. doi: 10.1037/a0033619. [SBr]

Dolan, R. J. & Dayan, P. (2013) Goals and habits in the brain. *Neuron* **80**(2):312–25. [arFC]

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D. & Daw, N. D. (2015) Model-based choices involve prospective neural activity. *Nature Neuroscience* **18**(5):767–72. [rFC]

Dowsett, E., Semmler, C., Bray, H., Ankeny, R. A. & Chur-Hansen, A. (2018) Neutralising the meat paradox: Cognitive dissonance, gender, and eating animals. *Appetite* **123**:280–88. [JQ-D]

Dreyfus, H. L. (2007) Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology* **20**(2):247–68. [PP]

Dunning, D., Leuenberger, A. & Sherman, D. A. (1995) A new look at motivated inference: Are self-serving theories of success a product of motivational forces? *Journal of Personality and Social Psychology* **59**:58–68. [WT]

Dutton, D. G. & Aron, A. P. (1974) Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology* **30**(4):510–17. [aFC]

Echterhoff, G., Higgins, E. T. & Levine, J. M. (2009) Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science* **4**(5):496–521. [RG]

Edelson, J., Alduncin, A., Krewson, C., Sieja, J. A. & Uscinski, J. E. (2017) The effect of conspiratorial thinking and motivated reasoning on belief in election fraud. *Political Research Quarterly* **70**:933–46. [JJVB]

Effron, D. A. (2018) It could have been true: How counterfactual thoughts reduce condemnation of falsehoods and increase political polarization. *Personality and Social Psychology Bulletin* **44**:729–45. [JJVB]

Egan, L. C., Bloom, P. & Santos, L. R. (2010) Choice-induced preferences in the absence of choice: Evidence from a blind two choice paradigm with young children and capuchin monkeys. *Journal of Experimental Social Psychology* **46**(1):204–207. [EH-J]

Egan, L. C., Santos, L. R. & Bloom, P. (2007) The origins of cognitive dissonance: Evidence from children and monkeys. *Psychological Science* **18**(11):978–83. [aFC]

Elliot, A. J. & Devine, P. G. (1994) On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology* **67**(3):382–94. [aFC]

Engel, C. & Glöckner, A. (2013) Role-induced bias in court: An experimental analysis. *Journal of Behavioral Decision Making* **26**(3):272–84. [DS]

Evans, J. S. B. (2019) Reflections on reflection: the nature and function of Type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*. (Advance online publication) [WDN]

Evans, J. S. B. & Wason, P. C. (1976) Rationalization in a reasoning task. *British Journal of Psychology* **67**:479–86. [WDN]

Evans, J. St. B. & Stanovich, K. E. (2013) Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science* **8**:223–41. [WDN]

Fazio, R. H. (1990) Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In: *Advances in Experimental Social Psychology, vol. 23*, ed. M. P. Zanna, pp. 75–109. Academic Press. [WT]

Fazio, R. H., Zanna, M. P. & Cooper, J. (1977) Dissonance and self-perception: An integrative view of each theory's proper domain of application. *Journal of Experimental Social Psychology* **13**:464–79. [EH-J]

Felin, T., Felin, M., Krueger, J. I. & Koenderink, J. (2019) On surprise-hacking. *Perception* **48**:109–14. [WT]

Festinger, L. & Carlsmith, J. M. (1959) Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology* **58**(2):203–10. [SA, TS]

Festinger, L. (1957) *A theory of cognitive dissonance*. Stanford University Press. [EH-J, DS]

Festinger, L. (1962) *A theory of cognitive dissonance, vol. 2*. Stanford University Press. [aFC, TS, WT]

Festinger, L. (1964) *Conflict, decision and dissonance*. Stanford University Press. [DS]

Festinger, L. (1999) Reflections on cognitive dissonance: 30 years later. In: *Cognitive dissonance: Progress on a pivotal theory in social psychology*, ed. E. Harmon-Jones & J. Mills, pp. 381–85. American Psychological Association. [aFC]

Festinger, L., Riecken, H. & Schacter, S. (1956) *When prophecy fails*. University of Minnesota Press. [aFC, NL]

Fishbein, M. & Ajzen, I. (1975) *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley. [WT]

Fodor, J. A. (1980) *The language of thought*. Harvard University Press. [ALR]

Foerde, K., Race, E., Verfaellie, M. & Shohamy, D. (2013) A role for the medial temporal lobe in feedback-driven learning: evidence from amnesia. *Journal of Neuroscience* **33**(13):5698–704. [aFC]

Foerde, K. & Shohamy, D. (2011) Feedback timing modulates brain systems for learning in humans. *Journal of Neuroscience* **31**(37):13157–67. [aFC]

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G. & Nosek, B.A. (2019) A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology* **117**(3):522–59. https://doi.org/10.1037/pspa0000160. [WT]

Frankish, K. (2004) *Mind and supermind*. Cambridge University Press. [WV]

Frederick, S. (2005) Cognitive reflection and decision making. *Journal of Economic Perspectives* **19**:24–42. [rFC]

Fredricks, A. J. & Dossett, D. L. (1983) Attitude-behavior relations: A comparison of the Fishbein-Ajzen and the Bentler-Speckart models. *Journal of Personality and Social Psychology* **45**:501–12. [WT]

Freeman, W. J. (1997) Nonlinear neurodynamics of intentionality. *Journal of Mind and Behavior* **18**(2/3):291–304. [PP]

Funkhouser, E. (2017) Beliefs as signals: A new function for belief. *Philosophical Psychology* **30**(6):809–31. Available at: https://doi.org/10.1080/09515089.2017.1291929 [NL]

Gawronski, B. & Strack, F. (2004) On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology* **40**(4):535–42. [SA]

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B. & Hu, X. (2018) Contextualized attitude change. In: *Advances in experimental social psychology, vol. 57*, ed. J. M. Olson, pp. 1–52. Elsevier Academic Press. [aFC]

Gazzaniga, M. (1985) *The social brain*. The Free Press. [WT]

Gazzaniga, M. S. (1967) The split brain in man. *Scientific American* **217**(2):24–29. [aFC]

Gendler, T. S. (2007) Self-deception as pretense. *Philosophical Perspectives* **21**(1):231–58. Available at: https://doi.org/10.1111/j.1520-8583.2007.00127.x. [JD]

Gendler, T. S. (2008) Alief in action (and reaction). *Mind & Language* **23**(5):552–85. [aFC]

Gershman, S. J. (2018) The successor representation: Its computational logic and neural substrates. *Journal of Neuroscience* **38**(33):7193–7200. [rFC]

Gershman, S. J. (2019) How to never be wrong. *Psychonomic Bulletin and Review* **26** (1):13–28. doi: https://doi.org/10.3758/s13423-018-1488-8. [rFC]

Gershman, S. J., Gerstenberg, T., Baker, C. L. & Cushman, F. A. (2016) Plans, habits, and theory of mind. *PLOS ONE* **11**(9):e0162246. [aFC, RS]

Gershman, S. J., Markman, A. B. & Otto, A. R. (2014) Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General* **143**(1):182–94. [arFC]

Gershman, S. J., Zhou, J. & Kommers, C. (2017) Imaginative reinforcement learning: Computational principles and neural mechanisms. *Journal of Cognitive Neuroscience* **29**(12):2103–113. [aFC]

Gibson, E. J. & Walk, R. D. (1960) The "visual cliff." *Scientific American* **202**(4):64–71. [aFC]

Gigerenzer, G. & Selten, R. (2002) *Bounded rationality: The adaptive toolbox*. MIT press. [aFC]

Gilbert, D.T. (2006) *Stumbling on happiness*. Vintage Books. [JQ-D]

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J. & Wheatley, T. P. (1998) Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology* **75**(3):617–38. [aFC]

Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. (2010) States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**:585–95. [rFC]

Glass, D.C. (1964) Changes in liking as a means of reducing cognitive discrepancies between self-esteem and aggression. *Journal of Personality* **32**(4):531–49. [JQ-D]

Glimcher, P. W. (2011) Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences USA* **108**(Suppl. 3):15647–54. [arFC]

Glöckner, A., Betsch, T. & Schindler, N. (2010) Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making* **23**:439–62. [DS]

Glöckner, A., Hilbig, B. E. & Jekel, M. (2014) What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition* **133**(3):641–66. [DS]

Gopnik, A. & Meltzoff, A. N. (1994) Minds, bodies, and persons: Young children's understanding of the self and others as reflected in imitation and theory of mind research. In: *Self-awareness in animals and humans: Developmental perspectives*, eds. S. T. Parker, R. W. Mitchell & M. L. Boccia, pp. 166–86. Cambridge University Press. [TDU]

Gopnik, A., Meltzoff, A. N. & Bryant, P. (1997) *Words, thoughts, and theories, vol. 1*. MIT Press. [aFC]

Graham, J. & Haidt, J. (2010) Beyond beliefs: Religions bind individuals into moral communities. *Personality and Social Psychology Review* **14**:140–50. [JG]

Graham, J. & Haidt, J. (2012) Sacred values and evil adversaries: A moral foundations approach. In: *The social psychology of morality: Exploring the causes of good and evil*, ed. P. Shaver & M. Mikulincer, pp. 11–31. APA Books. [JG]

Graham, J. & Yudkin, D. (in press) Variations in moral concerns across political ideology: Moral foundations, hidden tribes, and righteous division. In: *The Oxford Handbook of Moral Psychology*, ed. M. Vargas & J. Doris. Oxford University Press. [JG]

Graybiel, A. M. (2008) Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience* **31**:359–87. [aFC]

Greene, J. (2014) *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin. [aFC]

Greenwald, A. G. & Banaji, M. R. (1995) Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review* **102**(1):4–27. [aFC]

Griffiths, T. L., Lieder, F. & Goodman, N. D. (2015) Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science* **7**(2):217–29. [arFC, RS]

Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* **108**(4):814–34. [arFC, PP, SA]

Haidt, J. & Graham, J. (2009) The planet of the Durkheimians, where community, authority and sacredness are foundations of morality. In: *Social and psychological bases of ideology and system justification*, ed. J. T. Jost, A. C. Kay & H. Thorisdottir, pp. 371–401. Oxford University Press. [JG]

Hall, L., Johansson, P. & Strandberg, T. (2012) Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLOS ONE* **7**(9):e45457. Available at: https://doi.org/10.1371/journal.pone.0045457. [NL, PP]

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B. & Johansson, P. (2013) How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLOS ONE* **8**(4):e60554. [PP]

Harman, G. (1986) *Change in view: Principles of reasoning*. MIT Press. [AD]

Harmon-Jones, C., Schmeichel, B. J., Inzlicht, M. & Harmon-Jones, E. (2011) Trait approach motivation relates to dissonance reduction. *Social Psychological and Personality Science* **2**:21–28. [EH-J]

Harmon-Jones, E. (2000) Cognitive dissonance and experienced negative affect: Evidence that dissonance increases experienced negative affect even in the absence of aversive consequences. *Personality and Social Psychology Bulletin* **26**:1490–1501. [EH-J]

Harmon-Jones, E. (2019) *Cognitive dissonance: Re-examining a pivotal theory in psychology* (2nd ed.). American Psychological Association. [EH-J]

Harmon-Jones, E., Amodio, D. M. & Harmon-Jones, C. (2009) Action-based model of dissonance: A review, integration, and expansion of conceptions of cognitive conflict. In: *Advances in experimental social psychology, vol. 41*, ed. M. P. Zanna, pp. 119–66. Academic Press. [EH-J]

Harmon-Jones, E., Armstrong, J. & Olson, J. M. (2019) The influence of behavior on attitudes. In: *Handbook of attitudes, vol. 1: Basic principles* (2nd ed.), ed. D. Albarracin & B. T. Johnson, pp. 404–49). Routledge. [EH-J]

Harmon-Jones, E., Brehm, J. W., Greenberg, J., Simon, L. & Nelson, D. E. (1996) Evidence that the production of aversive consequences is not necessary to create cognitive dissonance. *Journal of Personality and Social Psychology* **70**:5–16. [aFC, EH-J]

Harmon-Jones, E. & Harmon-Jones, C. (2002) Testing the action-based model of cognitive dissonance: The effect of action-orientation on post-decisional attitudes. *Personality and Social Psychology Bulletin* **28**:711–23. [EH-J]

Harmon-Jones, E. & Harmon-Jones, C. (2019) Understanding the motivation underlying dissonance effects: The action-based model. In: *Cognitive Dissonance: Reexamining a pivotal theory in psychology* (2nd ed.), ed. E. Harmon-Jones, pp. 63–89. American Psychological Association. [EH-J]

Harmon-Jones, E., Harmon-Jones, C., Fearn, M., Sigelman, J. D. & Johnson, P. (2008) Left frontal cortical activation and spreading of alternatives: Tests of the action-based model of dissonance. *Journal of Personality and Social Psychology* **94**:1–15. doi: 10.1037/0022-3514.94.1.1. [EH-J]

Harmon-Jones, E., Harmon-Jones, C. & Levy, N. (2015a) An action-based model of cognitive dissonance processes. *Current Directions in Psychological Science* **24**:184–89. [EH-J]

Harmon-Jones, E. & Mills, J. (1999) An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In: *Science conference series. Cognitive dissonance: Progress on a pivotal theory in social psychology*, ed. E. Harmon-Jones & J. Mills, pp. 3–21. American Psychological Association. https://doi.org/10.1037/10318-001. [aFC]

Harmon-Jones, E., Price, T. F. & Harmon-Jones, C. (2015b) Supine body posture decreases rationalizations: Testing the action-based model of dissonance. *Journal of Experimental Social Psychology* **56**:228–34. [EH-J]

Hasselkus, B. R. (2011) *The meaning of everyday occupation*. Slack. [rFC]

Hawthorne-Madell, D. & Goodman, N. D. (2019) Reasoning about social sources to learn from actions and outcomes. *Decision* **6**(1):17–60. [rFC]

Hawton, K., Witt, K. G., Salisbury, T. L. T., Arensman, E., Gunnell, D., Hazell, P., Townsend, E. & van Heeringen, K. (2016) Psychosocial interventions following self-harm in adults: A systematic review and meta-analysis. *The Lancet Psychiatry* **3**(8): 740–50. [RS]

Hayek, F. A. (1973) *Law, legislation and liberty, vol. 1: Rules and order*. University of Chicago Press. [rFC]

Heider, F. (1979) On balance and attribution. In: *Perspectives on social network research*, ed. P. W. Holland & S. Leinhardt, pp. 11–23. Academic Press. [DS]

Heider, F. (1958/2013) *The psychology of interpersonal relations*. Psychology Press. [arFC]

Henrich, J. (2015) *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press. [NL]

Henrich, J. & Henrich, N. (2010) The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins. *Proceedings of the Royal Society B: Biological Sciences* **277**(1701):3715–24. Available at: https://doi.org/10.1098/rspb.2010.1191. [AD, aFC]

Hitler, A. (1925) *Mein kampf*. Verlag Franz Eher Nachfolger. [SBr]

Ho, M. K., MacGlashan, J., Greenwald, A., Littman, M. L., Hilliard, E. M., Trimbach, C., Brawner, S., Tenenbaum, J. B., Kleiman-Weiner, M. & Austerweil, J. L. (2016) Feature-based joint planning and norm learning in collaborative games. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, eds. A. Papafragou, D. J. Grodner, D. Mirman & J. Trueswell. Cognitive Science Society. https://mindmodeling.org/cogsci2016/index.html. [aFC]

Ho, M. K., MacGlashan, J., Littman, M. L. & Cushman, F. (2017) Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* **167**:91–106. [aFC]

Hodson, G., Dovidio, J. F. & Gaertner, S. L. (2002) Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin* **28**:460–71. [WT]

Holyoak, K. J. & Powell, D. (2016) Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin* **142**:1179–1203. [DS]

Holyoak, K. J. & Simon, D. (1999) Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General* **128**:3–31. [DS]

Holyoak, K. J. & Thagard, P. (1989) Analogical mapping by constraint satisfaction. *Cognitive Science* **13**:295–355. [DS]

Hume, D. (1738/1978) *A treatise of human nature*. Edited by L.A. Selby-Bigge & P. H. Nidditch. Oxford University Press. (Original work published in 1738) [PR]

Hume, D. (1739/2003) *A treatise of human nature*. Courier Corporation. [aFC]

Hunt, V., Layton, D. & Prince, S. (2015) *Diversity matters*. McKinsey. [WT]

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., S. J. Gershman, Dayan, P. & Roiser, J. P. (2015) Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences* **112**(10):3098–103. [aFC]

Icard, T., Cushman, F. & Knobe, J. (2018) On the instrumental value of hypothetical and counterfactual thought. In: *Proceedings of the 40th Annual Cognitive Science Society Meeting*, Madison, WI, pp. 517–22. Cognitive Science Society. https://mindmodeling.org/cogsci2018/index.html [aFC]

Inglehart, R. & Norris, P. (2003) *Rising tide: Gender equality and cultural change around the world*. Cambridge University Press. [WT]

Janis, I. L. & Mann, L. (1977) *Decision making: A psychological analysis of conflict, choice, and commitment*. Free Press. [DS]

Jimenez, A. (2015) *Psychosocial functioning and defenses among male pedophiles, adult sex offenders, and outpatients with depression and anxiety*. PsyD doctoral disseration, Adler School of Professional Psychology. UMI Order Number: AAI3664297. [SBr]

Johansson, P., Hall, L., Sikström, S. & Olsson, A. (2005) Failure to detect mismatches between intention and outcome in a simple decision task. *Science* **310**(5745):116–19. [aFC, PP]

Johansson, P., Hall, L., Tärning, B., Sikström, S. & Chater, N. (2014) Choice blindness and preference change: You will like this paper better if you (believe you) chose to read it! *Journal of Behavioral Decision Making* **27**(3):281–89. [NL, PP]

Jones, E. (1908) Rationalisation in every-day life. *The Journal of Abnormal Psychology* **3**(3):161–69. [JE]

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A. & Schoenbaum, G. (2012) Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338**(6109):953–56. [rFC]

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Jordan, J. J., Sommers, R., Bloom, P. & Rand, D. G. (2017) Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science* **28**(3):356–68. [SA]

Jost, J. T. (2006) The end of the end of ideology. *American Psychologist* **61**:651–70. [JG, rFC]

Jost, J. T. & Banaji, M. R. (1994) The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology* **33**(1):1–27. [aFC, JG, KL]

Jost, J. T., Banaji, M. R. & Nosek, B. A. (2004) A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology* **25**(6):881–919. [KL]

Jost, J. T., Becker, J., Osborne, D. & Badaan, V. (2017) Missing in (collective) action: Ideology, system justification, and the motivational antecedents of two types of protest behavior. *Current Directions in Psychological Science* **26**:99–108. [JG]

Jost, J. T., Glaser, J., Kruglanski, A. W. & Sulloway, F. J. (2003) Political conservatism as motivated social cognition. *Psychological Bulletin* **129**(3):339–75. [RG]

Jost, J. T., Ledgerwood, A. & Hardin, C. D. (2008) Shared reality, system justification, and the relational basis of ideological beliefs. *Social and Personality Psychology Compass* **2**(1):171–86. [RG]

Kahle, L. R. & Berman, J. J. (1979) Attitudes cause behaviors: A cross-lagged panel analysis. *Journal of Personality and Social Psychology* **37**:315–21. [WT]

Kahneman, D. (2011a) *Thinking, fast and slow*. Farrar, Straus and Giroux. [WDN]

Kahneman, D. (2011b) *Thinking, fast and slow*. Macmillan. [aFC]

Kaplan, J. T., Gimbel, S. I. & Harris, S. (2016) Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports* **6**:39589. [JJVB]

Kashdan, T. B., Ferssizidis, P., Collins, R. L. & Muraven, M. (2010) Emotion differentiation as resilience against excessive alcohol use: An ecological momentary assessment in underage social drinkers. *Psychological Science* **21**(9):1341–47. [RS]

Kay, A. C., Jimenez, M. C. & Jost, J. T. (2002) Sour grapes, sweet lemons, and the anticipatory rationalization of the status quo. *Personality and Social Psychology Bulletin* **28**(9):1300–12. [KL]

Kelleher, R. T. & Gollub, L. R. (1962) A review of positive conditioned reinforcement. *Journal of the Experimental Analysis of behavior* **5**(S4):543–97. [aFC]

Kelman, H. C. (1958) Compliance, identification, and internalization three processes of attitude change. *Journal of Conflict Resolution* **2**(1):51–60. [aFC]

Kennedy, K. A. & Pronin, E. (2008) When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin* **34**(6):833–48. [RS]

Keramati, M., Smittenaar, P., Dolan, R. J. & Dayan, P. (2016) Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences* **113**(45):12868–73. [aFC]

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L. & Tenenbaum, J. B. (2016) Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In: *Proceedings of the Cognitive Science Society*, ed. A. Papafragou, D. J. Grodner, D. Mirman & J. Trueswell. Cognitive Science Society. https://mindmodeling.org/cogsci2016/index.html [arFC]

Knoll, M., Starrs, C. J. & Perry, J. C. (2016) Rationalization (defense mechanism). In: *Encyclopedia of personality and individual differences*, ed. V. Zeigler-Hill & T. Shackelford, pp. 1–5. Springer. [SBr]

Knowlton, B. J., Mangels, J. A. & Squire, L. R. (1996) A neostriatal habit learning system in humans. *Science* **273**(5280):1399–1402. [aFC]

Knox, R. E. & Inkster, J. A. (1968) Postdecision dissonance at post time. *Journal of Personality and Social Psychology* **8**(4, Pt. 1):319–23. [aFC]

Kool, W., Cushman, F. A. & Gershman, S. J. (2018) Competition and cooperation between multiple reinforcement learning systems. In: *Goal-directed decision making: Computations and neural circuits*, ed. R. Morris, A. Bornstein & A. Shenhav, pp. 153–78. Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-812098-9.00007-3. [aFC]

Kool, W., Gershman, S. J. & Cushman, F. A. (2017) Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science* **28**(9):1321–33. [arFC]

Kostopoulou, O., Russo, J. E., Keenan, G., Delaney, B. C., Douiri, A. (2012) Information distortion in physicians' diagnostic judgments. *Medical Decision Making* **32**:831–40. doi:10.1177/0272989X12447241 [DS]

Krosnick, J. A. (1999) Survey research. *Annual Review of Psychology* **50**:537–67. [AD]

Krueger, J. I. & Funder, D. C. (2004) Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences* **27**(3):313–27. [AD]

Kruger, J. & Dunning, D. (1999) Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* **77**(6):1121–34. [RG]

Kunda, Z. (1990) The case for motivated reasoning. *Psychological Bulletin* **108**(3):480–98. [aFC, JE]

Kurzban, R. & Aktipis, A. (2007) Modularity and the social mind: Are psychologists too self-ish? *Personality and Social Psychology Review* **11**(2):131–49. [SA]

Landy, J. F. & Goodwin, G. P. (2015) Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science* **10**(4):518–36. [AD]

Langer, E. J. (1975) The illusion of control. *Journal of Personality and Social Psychology* **32**(2):311–28. https://doi.org/10.1037/0022-3514.32.2.311. [aFC]

Laurin, K. (2018) Inaugurating rationalization: Three field studies find increased rationalization when anticipated realities become current. *Psychological Science* **29**(4):483–95. [KL]

Laurin, K., Gaucher, D. & Kay, A. (2013a) Stability and the justification of social inequality. *European Journal of Social Psychology* **43**(4):246–54. [KL]

Laurin, K., Kay, A. C. & Fitzsimons, G. J. (2012) Reactance versus rationalization: Divergent responses to policies that constrain freedom. *Psychological Science* **23**(2): 205–209. [KL]

Laurin, K., Kay, A. C., Proudfoot, D. & Fitzsimons, G. J. (2013b) Response to restrictive policies: Reconciling system justification and psychological reactance. *Organizational Behavior and Human Decision Processes* **122**(2):152–62. [KL]

Lazarus, R. S. (1982) Thoughts on the relations between emotion and cognition. *American Psychologist* **37**(9):1019–24. https://doi.org/10.1037/0003-066X.37.9.1019. [aFC]

Leary, M. R. (1995) *Self-presentation: Impression management and interpersonal behavior*. Westview Press. [SA]

Levi, K. (1981) Becoming a hit man. *Urban Life* **10**(1):47–63. [SBr]

Levy, N. (2014) *Consciousness and moral responsibility*. Oxford University Press. [NL]

Levy, N. (2019a) Due deference to denialism: Explaining ordinary people's rejection of established scientific findings. *Synthese* **196**(1):313–27. [NL]

Levy, N. (2019b) Nudge, nudge, wink, wink: Nudging is giving reasons. *Ergo: An Open Access Journal of Philosophy* **6**(10). (Online publication) Available at: https://doi.org/10.3998/ergo.12405314.0006.010. [NL]

Levy, N. & Alfano, M. (2019) Knowledge from vice: Deeply social epistemology. *Mind*. Available at: https://ora.ox.ac.uk/objects/uuid:0c055497-9903-4c7a-b17e-63bd35d25c1d. https://doi.org/10.1093/mind/fzz017. [NL]

Lewis, D. (1969/2008) *Convention: A philosophical study*. Wiley. (Original work published in 1969) [PP]

Lieberman, D., Tooby, J. & Cosmides, L. (2007) The architecture of human kin detection. *Nature* **445**(7129):727–31. [aFC]

Lieberman, M. D., Ochsner, K. N., Gilbert, D. T. & Schacter, D. L. (2001) Do amnesiacs exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change. *Psychological Science* **12**(2):135–40. [aFC]

Lieder, F., Griffiths, T. L., Huys, Q. J. & Goodman, N. D. (2018) The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review* **25**(1): 322–49. [RG]

Lifton, R. J. (1986) *The Nazi doctors: Medical killing and the psychology of genocide*. Basic Books. [SBr]

Liu, B. S. & Ditto, P. H. (2012) What dilemma? Moral evaluation shapes factual belief. *Social Psychological and Personality Science* **4**(3):316–23. [PKS]

Lodder, P., Ong, H. H., Grasman, R. P. P. P. & Wicherts, J. (2019) A comprehensive meta-analysis of money priming. *Journal of Experimental Psychology: General* **148**(4):688–712. https://doi.org/10.1037/xge0000570. [WT]

Lombrozo, T. (2017) "Learning by thinking" in science and everyday life. In: *The scientific imagination*. Oxford University Press. [aFC, TDU]

Lord, C. G., Ross, L. & Lepper, M. R. (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* **37**(11):2098–2109. [RG]

Losch, M. E. & Cacioppo, J. T. (1990) Cognitive dissonance may enhance sympathetic tonus, but attitudes are changed to reduce negative affect rather than arousal. *Journal of Experimental Social Psychology* **26**(4):289–304. [aFC]

Loughnan, S., Haslam, N. & Bastian, B. (2010) The role of meat consumption in the denial of moral status and mind to meat animals. *Appetite* **55**:156–59. [JQ-D]

Luo, J. & Yu, R. (2017) The spreading of alternatives: Is it the perceived choice or actual choice that changes our preference? *Journal of Behavioral Decision Making* **30**(2):484–91. [PP]

Mach, E. (1897/1976) On thought experiments. In: *Knowledge and error* (6th ed.), by E. Mach, trans. T. McCormack & P. Foulkes, pp. 134–47. D. Reidel. (Original work published in 1897) [TDU]

Mandelbaum, E. (2019) Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language* **34**:141–57. [JQ-D]

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman. [arFC]

Mazzarella, D., Reinecke, R., Noveck, I. & Mercier, H. (2018) Saying, presupposing and implicating: How pragmatics modulates commitment. *Journal of Pragmatics* **133**:15–27. [SA]

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., Acar, O. A., Aczel, B., Bakos, B. E., F. Barbosa, F., Baskin, L. Bègue, E., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., Coary, S. P., Crusius, J., Evans, J. R., Feldman, N., Ferreira-Santos, F., Gamer, M., Gerlsma, C., Gomes, S., González-Iraizoz, M., Holzmeister, F., Huber, J., Huntjens, R. J. C., Isoni, A., Jessup, R. K., Kirchler, M., klein Selle, N., Koppel, L., Kovacs, M., Laine, T., Lentz, F., Loschelder, D. D., Ludvig, E. A., Lynn, M. L., Martin, S. D., McLatchie, N. M., Mechtel, M., Nahari, G., Özdoğru, A. A., Pasion, R., Pennington, C. R., Roets, A., Rozmann, N., Scopelliti, I., Spiegelman, E., Suchotzki, K., Sutan, A., Szecsi, P., Tinghög, G., Tisserand, J.-C., Tran, U. S., Van Hiel, A., Vanpaemel, W., Västfjäll, D., Verliefde, T., Vezirian, K., Voracek, M., Warmelink, L., Wick, K., Wiggins, B. J.,

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press

Wylie, K. & Yildiz, E. (2018) Registered replication report on Srull & Wyer (1979) *Advances in Methods and Practices in Psychological Science* **1**:321–336. [WT]

McClelland, J. L. & Rumelhart, D. E. (1981) An interactive activation model of context effects in letter perception: Part I: An account of basic findings. *Psychological Review* **88**:375–407. [DS]

McCoy, J., Paul, L. & Ullman, T. (2019) Modal prospection. In: *Metaphysics and cognitive science*, ed. A. I. Goldman & B. P. McLaughlin, pp. 235–67. Oxford University Press. [rFC, TDU]

Mele, A. (2007) Persisting intentions. *Noûs* **41**:735–57. [SBe]

Menon, M., Tobin, D. D., Corby, B. C., Hodges, E. V. & Perry, D. G. (2007) The developmental costs of high self-esteem for antisocial children. *Child Development* **78**(6): 1627–39. doi: 10.1111/j.1467-8624.2007.01089.x. [SBr]

Mercier, H. & Sperber, D. (2011) Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* **34**(2):57–74. [arFC, PP, RS, SA]

Mercier, H. (2012) The social functions of explicit coherence evaluation. *Mind & Society* **11**(1):81–92. [SA]

Mercier, H. & Sperber, D. (2017) *The enigma of reason*. Harvard University Press. [SA, WDN]

Miller, G. A. & Buckhout, R. (1962/1973) *Psychology: The science of mental life*. Harper & Row. [aFC]

Miller, K. J., Shenhav, A. & Ludvig, E. A. (2019) Habits without values. *Psychological Review* **126**(2):292–311. https://doi.org/10.1037/rev0000120. [rFC]

Monteith, M. J., Devine, P. G. & Zuwerink, J. R. (1993) Self-directed versus other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology* **64**:198–210. [WT]

Mook, D. G. (1983) In defense of external invalidity. *American Psychologist* **38**(4):379–87. [rFC]

Moore, M. T. & Fresco, D. M. (2012) Depressive realism: A meta-analytic review. *Clinical Psychological Review* **32**:496–509. [JQ-D]

Morris, A. & Cushman, F. (2017) A common framework for theories of norm compliance. *Social Philosophy and Policy* **35**(1):101–27. [aFC]

Morris, A. & Cushman, F. (2019) Model-free RL or action sequences? *Frontiers in Psychology* **10**. https://doi.org/10.3389/fpsyg.2019.02892 [rFC]

Morris, A., Phillips, J. S., Icard, T., Knobe, J., Gerstenberg, T. & Cushman, F. A. (2018, April 26). Causal judgments approximate the effectiveness of future interventions. *PsyArxiv*: https://doi.org/10.31234/osf.io/nq53z. [aFC]

Morris, G., Nevet, A., Arkadir, D., Vaadia, E. & Bergman, H. (2006) Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience* **9**(8):1057–63. [aFC]

Neidigh, L. & Krop, H. (2015) Cognitive distortions among child sexual offenders. *Journal of Sex Education and Therapy* **18**(3):208–15. doi: 10.1080/01614576.1992.11074054. [SBr]

Neisser, U. (2014) *Cognitive psychology: Classic edition*. Psychology Press. [aFC]

Nesse, R. M. & Ellsworth, P. C. (2009) Evolution, emotions, and emotional disorders. *American Psychologist* **64**:129–39. [PR]

Newman, I., Gibb, M. & Thompson, V. A. (2017) Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **43**:1154–70. [WDN]

Ng, A. Y. & Russell, S. J. (2000) Algorithms for inverse reinforcement learning. In: *International Conference on Machine Learning*, ed. P. Langley, pp. 663–70. Morgan Kaufmann Publishers. [aFC]

Nickerson, R. S. (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* **2**(2):175–220. [aFC]

Nisbett, R. E. & Wilson, T. D. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**(3):231–59. Available at: https://doi.org/10.1037/0033-295X.84.3.231. [arFC, NL, SA, TS]

Nolen-Hoeksema, S. (2000) The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology* **109**(3):504–11. [RS]

Norman, D. A. & Shallice, T. (1986) Attention to action. In: *Consciousness and self-regulation: Advances in research and theory IV*, ed., R. Davidson, R. Schwartz & D. Shapiro, pp. 1–18. Plenum Press. [aFC]

Norton, M. I., Vandello, J. A. & Darley, J. M. (2004) Casuistry and social category bias. *Journal of Personality and Social Psychology* **87**:817–31. [WT]

O'Donnell, T. J. (2015) *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press. [aFC]

Okasha, S. (2018) *Agents and goals in evolution*. Oxford University Press. [WV]

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J. & Tetlock, P. E. (2015) Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology* **108**(4):562–71. [WT]

Pennycook, G., Fugelsang, J. A. & Koehler, D. J. (2015) What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology* **80**:34–72. [WDN]

Pereira, A., Harris, E. A. & Van Bavel, J. J. (2019) Identity concerns drive belief in fake news. (Unpublished manuscript). Available at: https://psyarxiv.com/7vc5d/. [JJVB]

Pessoa, L. (2008) On the relationship between emotion and cognition. *Nature Reviews Neuroscience* **9**:148–58. [PR]

Phillips, J. & Cushman, F. (2017) Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences* **114**(18):4649–54. [aFC]

Piazza, J., Ruby, M. B., Loughnan, S., Luong, M., Kulik, J., Watkins, H. M. & Seigerman, M. (2015) Rationalizing meat consumption: The 4Ns. *Appetite* **91**:114–28. [JQ-D]

Pinker, S. (2018) *Enlightenment now: The case for reason, science, humanism, and progress*. Penguin. [rFC]

Pizarro, D. A., Inbar, Y. & Helion, C. (2011) On disgust and moral judgment. *Emotion Review* **3**(3):267–68. Available at: https://doi.org/10.1177/1754073911402394. [AD]

Pizarro, D. A. & Uhlmann, E. L. (2005) Do normative standards advance our understanding of moral judgment? *Behavioral and Brain Sciences* **28**:558–59. [WT]

Pond Jr, R. S., Kashdan, T. B., DeWall, C. N., Savostyanova, A., Lambert, N. M. & Fincham, F. D. (2012) Emotion differentiation moderates aggressive tendencies in angry people: A daily diary analysis. *Emotion* **12**(2):326–37. [RS]

Price, J. P. (2007) Cognitive schemas, defence mechanisms and post-traumatic stress symptomatology. *Psychology and Psychotherapy* **80**(3):343–53. doi: 10.1348/147608306X144178. [SBr]

Pyszczynski, T., Solomon, S. & Greenberg, J. (2015) Thirty years of terror management theory: From genesis to revelation. *Advances in Experimental Social Psychology* **52**:1–70. [JQ-D]

Rabin, M. & Schrag, J. (1999) First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics* **114**(1):37–82. [SEW]

Railton, P. (2014) The affective dog and its rational tale: Intuition and attunement. *Ethics* **124**(4):813–59. [aFC]

Ramsey, F. P. (1931) *The foundations of mathematics and other essays*. Routledge and Kegan Paul. [JD]

Randall, D. M. & Wolff, J. A. (1994) The time interval in the intention-behaviour relationship: Meta-analysis. *British Journal of Social Psychology* **33**:405–18. [WT]

Rawls, J. (1971) *A theory of justice*. Harvard university press. [PP]

Read, S. J. & Marcus-Newhall, A. (1993) Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology* **65**(3):429–47. [aFC]

Read, S. J. & Simon, D. (2012) Parallel constraint satisfaction as a mechanism for cognitive consistency. In: *Cognitive consistency: A fundamental principle in social cognition*, ed. B. Gawronski & F. Strack, pp. 66–86. Guilford Press. [DS]

Read, S. J., Vanman, E. J. & Miller, L. C. (1997) Connectionism, parallel constraint satisfaction processes, and gestalt principles: (Re)introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review* **1**:26–53. [DS]

Reker, G. T., Peacock, E. J. & Wong, P. T. (1987) Meaning and purpose in life and well-being: A life-span perspective. *Journal of Gerontology* **42**(1):44–49. [rFC]

Richerson, P. J. & Boyd, R. (2008) *Not by genes alone: How culture transformed human evolution*. University of Chicago Press. [aFC, NL]

Roesch, M. R., Calu, D. J. & Schoenbaum, G. (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience* **10**(12):1615–24. [aFC]

Rothgerber, H. (2014) Efforts to reduce vegetarian-induced dissonance among meat eaters. *Appetite* **79**:32–41. [JQ-D]

Royzman, E. & Hagan, J. P. (2017) The shadow and the tree: Inference and transformation of cognitive content in psychology of moral judgment. In: *Moral inferences*, ed. J.-F. Bonefon & B. Trémolière, pp. 56–74. Taylor & Francis. [AD]

Rozin, P., Hormes, J. M., Faith, M. S. & Wansink, B. (2012) Is meat male? A quantitative multimethod framework to establish metaphoric relationships. *Journal of Consumer Research* **39**(3):629–43. [JQ-D]

Rush Burkey, C. & ten Bensel, T. (2015) An examination and comparison of rationalizations employed by solo and co-offending female sex offenders. *Violence and Gender* **2**(3): 168–78. doi: 10.1089/vio.2015.0018. [SBr]

Russo, J. E., Carlson, K. A., Meloy, M. G. & Yong, K. (2008) The goal of consistency as a cause of information distortion. *Journal of Experimental Psychology: General* **137** (3):456–70. [DS]

Russo, J. E., Medvec, V. H. & Meloy, M. G. (1996) The distortion of information during decisions. *Organizational Behavior and Human Decision Processes* **66**(1): 102–10. [DS]

Russo, J. E., Meloy, M. G. & Medvec, V. H. (1998) Predecisional distortion of product information. *Journal of Marketing Research* **35**(4):438–52. [DS]

Saxe, R. (2005) Against simulation: The argument from error. *Trends in Cognitive Sciences* **9**:174–79. [rFC]

Saxe, R. (2009) The happiness of the fish: Evidence for a common theory of one's own and others' actions. In: *The handbook of imagination and mental simulation*, ed. K. D. Markman, W. M. P. Klein & J. A. Suhr, pp. 257–66. Psychology Press. [TDU]

Schachter, S. & Singer, J. (1962) Cognitive, social, and physiological determinants of emotional state. *Psychological Review* **69**(5):379–99. https://doi.org/10.1037/h0046234b [aFC]

Schultz, W., Dayan, P. & Montague, P. R. (1997) A neural substrate of prediction and reward. *Science* **275**(5306):1593–99. [rFC]

Schwitzgebel, E. (2002) A phenomenal, dispositional account of belief. *Noûs* **36**(2):249–75. Available at: https://doi.org/10.1111/1468-0068.00370. [NL]

Schwitzgebel, E. & Ellis, J. (2017) Rationalization in moral and philosophical thought. In: *Current issues in thinking and reasoning. Moral inferences*, ed. J.-F. Bonnefon & B. Trémolière, pp. 170–90. Routledge/Taylor & Francis. [JE]

Shafir, E., Simonson, I. & Tversky, A. (1993) Reason-based choice. *Cognition* **49**(1–2):11–36. [SA]

Shalvi, S., Gino, F., Barkan, R. & Ayal, S. (2015) Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science* **24**(2):125–30. Available at: https://doi.org/10.1177/0963721414553264. [JD]

Shanteau, J. (1992) Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes* **53**:252–66. [SEW]

Sharot, T., De Martino, B. & Dolan, R. J. (2009) How choice reveals and shapes expected hedonic outcome. *Journal of Neuroscience* **29**(12):3760–65. [aFC]

Sharot, T., Velasquez, C. M. & Dolan, R. J. (2010) Do decisions shape preference? Evidence from blind choice. *Psychological Science* **21**(9):1231–35. Available at: https://doi.org/10.1177/0956797610379235. [AD, arFC, TS]

Shenhav, A., Botvinick, M. M. & Cohen, J. D. (2013) The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron* **79**(2):217–40. [arFC]

Sheppard, B. H., Hartwick, J. & Warshaw, P. R. (1988) The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research* **15**:325–43. [WT]

Shultz, T. R. & Lepper, M. R. (1999) Computer simulation of cognitive dissonance reduction. In: *Cognitive dissonance: Progress on a pivotal theory in social psychology*, ed. E. Harmon-Jones & J. Mills, pp. 235–65. American Psychological Association. https://doi.org/10.1037/10318-010. [aFC]

Simon, D. (2004) A third view of the black box: Cognitive coherence in legal decision making. *University of Chicago Law Review* **71**:511–86. [DS]

Simon, D. & Holyoak, K. J. (2002) Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and Social Psychology Review* **6**:283–94. [DS]

Simon, D., Krawczyk, D. C. & Holyoak, K. J. (2004a) Construction of preferences by constraint satisfaction. *Psychological Science* **15**:331–36. [DS]

Simon, D., Pham, L. B., Le, Q. A. & Holyoak, K. J. (2001) The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **27**:1250–60. [DS]

Simon, D., Snow, C. J. & Read, S. J. (2004b) The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology* **86**:814–37. [DS]

Simon, D. & Spiller, S. A. (2016) The elasticity of preferences. *Psychological Science* **27**:1588–99. [DS]

Simon, D., Stenstrom, D. M. & Read, S. J. (2015) The coherence effect: Blending cold and hot cognitions. *Journal of Personality and Social Psychology* **109**:369–94. [DS]

Simon, D. A. & Daw, N. D. (2011) Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience* **31**(14):5526–39. [rFC]

Simonson, I. (1989) Choice based on reasons: The case of attraction and compromise effects. *The Journal of Consumer Research* **16**(2):158–74. [SA]

Sloman, S. A. (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* **119**(1):3–22. [aFC]

Sloman, S. A. & Rabb, N. (2016) Your understanding is my understanding: Evidence for a community of knowledge. *Psychological Science* **27**(11):1451–60. [rFC, RG]

Smith, K. S. & Graybiel, A. M. (2016) Habit formation coincides with shifts in reinforcement representations in sensorimotor striatum. *Journal of Neurophysiology* **115**:1487–98. [PR]

Spellman, B. A., Ullman, J. B. & Holyoak, K. J. (1993) A coherence model of cognitive consistency. *Journal of Social Issues* **4**:147–65. [DS]

Spencer, J. P., Blumberg, M. S., McMurray, B., Robinson, S. R., Samuelson, L. K. & Tomblin, J. B. (2009) Short arms and talking eggs: Why we should no longer abide the nativist–empiricist debate. *Child Development Perspectives* **3**(2):79–87. Available at: https://doi.org/10.1111/j.1750-8606.2009.00081.x. [AD]

Sperber, D. (2000) Metarepresentations in an evolutionary perspective. In: *Metarepresentations: A multidisciplinary perspective*, ed. D. Sperber, pp. 117–37). Oxford University Press. [SA]

Squire, L. R. (2004) Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory* **82**(3):171–77. [aFC]

Stafford, T. (2014) The perspectival shift: how experiments on unconscious processing don't justify the claims made for them. *Frontiers in Psychology* **5**: article no. 1067. doi:10.3389/fpsyg.2014.01067. Available at: https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01067/full. [TS]

Stafford, T. (2015) *For argument's sake: Evidence that reason can change minds.* (Self-published ebook) [TS]

Stanford, P. K. (2018) The difference between ice cream and Nazis: Moral externalization and the evolution of human cooperation. *Behavioral and Brain Sciences* **41**:E95. doi: 10.1017/S0140525X17001911. [PKS]

Stauffer, W. R., Lak, A. & Schultz, W. (2014) Dopamine reward prediction error responses reflect marginal utility. *Current Biology* **24**:2419–2500. [PR]

Steele, C. M. (1988) The psychology of self-affirmation: Sustaining the integrity of the self. In: *Advances in experimental social psychology, vol. 21, Social psychological studies of the self: Perspectives and programs*, ed. L. Berkowitz, pp. 261–302. Academic Press. [aFC]

Steele, C. M., Spencer, S. J. & Lynch, M. (1993) Self-image resilience and dissonance: The role of affirmational resources. *Journal of Personality and Social Psychology* **64**(6):885–96. [aFC]

Strandberg, T., Sivén, D., Hall, L., Johansson, P. & Pärnamets, P. (2018) False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology: General* **147**(9):1382–99. [PP]

Suppes, A., Napier, J. L., van der Toorn, J. (2019) The palliative effects of system justification on the health and happiness of lesbian, gay, bisexual, and transgender individuals. *Personality and Social Psychology Bulletin* **45**(3):372–88. [KL]

Sutton, R. S. (1991) Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin* **2**(4):160–63. [aFC]

Sutton, R. S. & Barto, A. G. (1998) *Reinforcement learning: An introduction, vol. 1.* MIT Press. [aFC]

Sutton, R. S., Precup, D. & Singh, S. (1999) Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* **112** (1–2):181–211. [aFC]

Sykes, G. M. & Matza, D. (1957) Techniques of neutralization: A theory of delinquency. *American Sociological Review* **22**(6):664–70. [SBr]

Tannenbaum, D., Valasek, C. J., Knowles, E. D. & Ditto, P. H. (2013) Incentivizing wellness in the workplace: Sticks (not carrots) send stigmatizing signals. *Psychological Science* **24**:1512–22. [WT]

Tarnita, C. E. (2017) The ecology and evolution of social behavior in microbes. *Journal of Experimental Biology* **220**:18–24. [WV]

Taylor, S. E. & Brown, J. D. (1988) Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin* **103**:193–210. [WT]

Taylor, W. S. (1923) Rationalization and its social significance. *Journal of Abnormal Psychology and Social Psychology* **17**(4):410–18. [SBr]

Tedeschi, J. T. & Rosenfeld, P. (1981) Impression management theory and the forced compliance situation. In: *Impression management theory and social psychological research*, ed. J. T. Tedeschi, pp. 147–77. [SA]

Tedeschi, J. T., Schlenker, B. R. & Bonoma, T. V. (1971) Cognitive dissonance: Private ratiocination or public spectacle. *American Psychologist* **26**(8):685–95. [arFC, SA]

Tennie, C., Call, J. & Tomasello, M. (2009) Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**(1528):2405–15. Available at: https://doi.org/10.1098/rstb.2009.0052. [NL]

Tetlock, P. E. (1992) The impact of accountability on judgment and choice : Toward a social contingency model. In: *Advances in experimental social psychology, vol. 25*, ed. M. P. Zanna, p. 331–76. Academic Press. https://psycnet.apa.org/doi/10.1016/S0065-2601(08)60287-7. [SA]

Tetlock, P. E. (2002) Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review* **109**(3):451–71. [arFC]

Thagard, P. (1989) Explanatory coherence. *Behavioral and Brain Sciences* **12**(3):435–67. [aFC, DS]

Thomas, A. J., Stanford, P. K. & Sarnecka, B. W. (2016) No child left alone: Moral judgments about parents affect estimates of risk to children. *Collabra* **2**:10. doi: http://doi.org/10.1525/collabra.33. [PKS]

Thorndike, E. L. (1898) Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review Monograph Supplements* **2**(4):i–109. [aFC]

Thornton, M. A. & Tamir, D. I. (2017) Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences USA* **114**:5982–87. [PR]

Tice, D. M. (1992) Self-concept change and self-presentation: The looking glass self is also a magnifying glass. *Journal of Personality and Social Psychology* **63**(3):435–51. [SA]

Tinbergen, N. (1963) On aims and methods of ethology. *Zeitschrift für Tierpsychologie* **20** (4):410–33. [arFC]

Tomasello, M. (2005) Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* **28**:675–35. [rFC]

Tomasello, M. (2014) *A natural history of human thinking.* Harvard University Press. [arFC, NL]

Tomasello, M. (in press) The moral psychology of obligation. *Behavioral and Brain Sciences.* [PKS]

Tomasello, M., Davis-Dasilva, M., Camak, L. & Bard, K. (1987) Observational learning of tool-use by young chimpanzees. *Human Evolution* **2**(2):175–83. [aFC]

Tooby, J. and Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind: Evolutionary psychology and the generation of culture*, ed. J. Barkow, L. Cosmides & J. Tooby, pp. 19–136. Oxford University Press. [PP]

Trapnell, P. D. & Campbell, J. D. (1999) Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology* **76**(2):284–304. [RS]

Trivers, R. (2000) The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences* **907**(1):114–31. [aFC, PP, RS]

Turiel, E. (2010) Snap judgment? Not so fast: Thought, reasoning, and choice as psychological realities. *Human Development* **53**(3):105–109. Available at: https://doi.org/10.1159/000315167. [AD]

Turiel, E. & Dahl, A. (2019) The development of domains of moral and conventional norms, coordination in decision-making, and the implications of social opposition. In: *The normative animal: On the biological significance of social, moral, and linguistic norms*, ed. K. Bayertz & N. Boughley, pp. 195–213. Oxford University Press. [AD]

Turner, J. C., Oakes, P. J., Haslam, S. A. & McGarty, C. (1994) Self and collective: Cognition and social context. *Personality and Social Psychology Bulletin* **20**:454–63. [JJVB]

Uhlmann, E. L. & Cohen, G. L. (2005) Constructed criteria: Redefining merit to justify discrimination. *Psychological Science* **16**:474–80. [WT]

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D. & Ditto, P. H. (2009) The motivated use of moral principles. *Judgment and Decision Making* **4**:476–91. [WT]

Ullman, T. D., Siegel, M., Tenenbaum, J. B. & Gershman, S. J. (2016) Coalescing the vapors of human experience into a viable and meaningful comprehension. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, ed. A. Papafragou, Daniel J. Grodner, D. Mirman & J. Trueswell. Cognitive Science Society. https://mindmodeling.org/cogsci2016/index.html. [rFC]

Uscinski, J. E. & Parent, J. M. (2014) *American conspiracy theories*. Oxford University Press. [JJVB]

Van Bavel, J. J. & Pereira, A. (2018) The partisan brain: An Identity-based model of political belief. *Trends in Cognitive Sciences* **22**(3):213–24. [JJVB, RG]

van Veen, V., Krug, M. K., Schooler, J. W. & Carter, C. S. (2009) Neural activity predicts attitude change in cognitive dissonance. *Nature Neuroscience* **12**(11):1469–74. [aFC]

Veit, W. (2019) Evolution of multicellularity: Cheating done right. *Biology & Philosophy* **34**:34. (Online first). Available at: https://doi.org/10.1007/s10539-019-9688-9. [WV]

Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N. & Daw, N. D. (2019) Hippocampal contributions to model-based planning and spatial memory. *Neuron* **102**(3):683–93. [rFC]

Vinckier, F., Rigoux, L., Kurniawan, I. T., Hu, C., Bourgeois-Gironde, S., Daunizeau, J. & Pessiglione, M. (2019) Sour grapes and sweet victories: How actions shape preferences. *PLOS Computational Biology* **15**(1):e1006499. [aFC, TS]

Von Hippel, W. & Trivers, R. (2011) Reflections on self-deception. *Behavioral and Brain Sciences* **34**(1):41–56. [arFC]

Von Uexküll, J. (1934/2010) *A foray into the worlds of animals and humans: With a theory of meaning*. University of Minnesota Press. (Original work published in 1934) [PP]

Vullioud, C., Clément, F., Scott-Phillips, T. & Mercier, H. (2017) Confidence as an expression of commitment: Why misplaced expressions of confidence backfire. *Evolution and Human Behavior* **38**(1):9–17. [SA]

Ward, L. C. & Rothaus, P. (1991) The measurement of denial and rationalization in male alcoholics. *Journal of Clinical Psychology* **47**(3):465–68. [SBr]

Wason, P. C. & Evans, J. S. B. (1975) Dual processes in reasoning? *Cognition* **3**:141–54. [WDN]

Webb, T. L. & Sheeran, P. (2006) Does changing behavioural intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin* **132**:249–68. [WT]

Webster, D. M. & Kruglanski, A. W. (1994) Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology* **67**(6):1049–62. [rFC]

Wheatley, T. & Haidt, J. (2005) Hypnotic disgust makes moral judgments more severe. *Psychological Science* **16**(10):780–84. [AD]

Whiten, A., McGuigan, N., Marshall-Pescini, S. & Hopper, L. M. (2009) Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**(1528):2417–28. [aFC]

Wike, R. & Castillo, A. (2018) *Many around the world are disengaged from politics* [Report]. Pew Research Center. Available at: https://www.pewresearch.org/global/2018/10/17/international-political-engagement/. [KL]

Wilson, T. D. (2004) *Strangers to ourselves*. Harvard University Press. [aFC]

Wilson, T. D. & Gilbert, D. T. (2003) Affective forecasting. *Advances in Experimental Social Psychology* **35**:345–411. [PR]

Wilson, T. D. C. & Nisbett, R. E. (1978) The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology* **41**(2):118–31. [TS]

Winthrop, R. C. (1852) *Addresses and speeches on various occasions: 1835–1851, vol. 1*. Little, Brown. [JG]

Wittgenstein, L. (1953/2009) *Philosophical investigations*. Wiley. (Original work published in 1953) [PP]

Wombacher, K., Matig, J. J., Sheff, S. E. & Scott, A. M. (2019) "It just kind of happens": College students' rationalizations for blackout drinking. *Health Communication* **34**(1): 1–10. doi: 10.1080/10410236.2017.1384351. [SBr]

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. (2010) Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**:686–88. [WT]

World Health Organization. (1993) *The ICD-10 classification of mental and behavioural disorders. Diagnostic criteria for research*. World Health Organization. [SBr]

Zaki, L. F., Coifman, K. G., Rafaeli, E., Berenson, K. R. & Downey, G. (2013) Emotion differentiation as a protective factor against nonsuicidal self-injury in borderline personality disorder. *Behavior Therapy* **44**(3):529–40. [RS]

Zanna, M. P. & Cooper, J. (1974) Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology* **29**(5):703–709. [aFC]

Zentall, T. R. & Singer, R. A. (2007) Within-trial contrast: Pigeons prefer conditioned reinforcers that follow a relatively more rather than less aversive event. *Journal of the Experimental Analysis of Behavior* **88**:131–149. [EH-J]

https://doi.org/10.1017/S0140525X19001730 Published online by Cambridge University Press