UNCONSCIOUS RATIONALIZATION, OR: *HOW (NOT) TO THINK ABOUT AWFULNESS AND DEATH*

Jake Quilty-Dunn (WUSTL)
Unpublished draft—March 2021

"How strange it is that man on earth should roam,
and lead a life of woe, but not forsake
his rugged path; nor dare he view alone
his future doom which is but to awake."
—Keats, "On Death"

1. Consciousness and human nature

Debates between empiricists and rationalists are often taken to rest on two major points of contention (e.g., Markie 2017): whether knowledge is ultimately grounded in perception alone or (at least partly) in *a priori* sources, and whether mental contents are acquired through perception alone or are (at least partly) innate. Another point of contention, however, concerns the process of thinking. Is thinking fundamentally a matter of triggering arational associative links, or is there irreducible logical structure? For rationalists, logical structure is required to distinguish an associatively generated string of representations like BODIES; HEAVY from the thought that bodies are heavy (Kant 1781/1787, B140–B143; Fodor 2003, 13), as well as to distinguish logical inference from other sorts of mental transitions (Fodor & Pylyshyn 1988).

In contemporary philosophy, the rationalists have made significant headway in the last of these disputes: there are few thoroughgoing Humeans who would reduce all cognition to association—not even neo-empiricists (Prinz 2002). The debate about associationism has not ended, however, but has merely been confined to the unconscious. Associationism thrives as a thesis about *unconscious* cognition. On one prominent view, rational, logical thought is exclusively—or at least primarily—conscious. Consciousness, according to proponents of this view, is constitutive of logical inference. Without it, cognitive processing may be useful and even broadly truth-tracking but fails to count as rational or guided by logical rules such as *modus ponens*. Call the view on which unconscious cognition is primarily associative *unconscious associationism*.

Philosophers are often motivated to endorse unconscious associationism because of a family of assumptions: genuine inference reflects on the rationality of the thinker; only what the thinker is responsible for can reflect on her rationality; the thinker can only be responsible for what she is conscious of. Paul Boghossian (2014; 2018) provides a usefully clear defense of this sort of perspective, which is generally shared by epistemological internalists (Neta 2013; Dogramaci 2014; Valaris 2014; Malmgren 2018; cf. Siegel 2017).

In cognitive science, unconscious associationism prevails through so-called "dual-process" theories of cognition (Kahneman 2011; Evans & Stanovich 2013). On dual-process theories, mental processing breaks down into two types: type-1 cognition, which is quick, automatic, unconscious, and associative, and type-2 cognition, which is slow, effortful, conscious, and logical. Dual-process theorists hold that logical inference requires the conscious use of working memory resources to guide transitions in thought.

For both the epistemological internalists and the dual-process theorists, unconscious cognition must be in some broad sense arational, and association is the most salient form of arational thought. Some proponents of unconscious associationism allow that there may be episodes of unconscious reasoning, but only derivatively: in such cases, a feature of conscious reasoning has "gone tacit" (Boghossian 2016, 48) and fails to conform to the "Platonic Form of reasoning" (*Ibid.*), i.e., conscious reflection.

Another tradition takes consciousness to be relatively unimportant for logical reasoning. On this view, the logical, rational aspects of human cognition lie much deeper than the glimpses we catch when looking inward. Call the view on which much of unconscious cognition has logical form *deep rationalism*. Deep rationalism may have reached a high point during the heyday of Freudian psychoanalytic theory, but it persists in many other forms. Perhaps the most representative defender of deep rationalism today is Noam Chomsky (1965; 2016). For Chomsky, consciousness reveals only "scattered fragments" (2016, 14) of thought, which consists primarily in "internal processes of mental computation that are unconscious, often inaccessible to consciousness, and very likely a core feature of fundamental human nature" (2018, 45).

As Chomsky points out, what is ultimately at stake in the dispute between unconscious associationism and deep rationalism is our picture of human nature. Are the richly structured, logical properties that characterize human thought dependent upon consciousness, or do they extend far deeper into the mind? Nicholas Chater has recently argued that "the mind is flat": "Our flow of momentary consciousness is not the sparkling surface of a vast sea of thought—it is *all there is*" (2018, 8).

This paper argues that the mind is not flat. I'll push instead for a particular, heterodox brand of deep rationalism. While deep rationalism is already quite heterodox, this paper will focus on

*ir*rational aspects of unconscious cognition. I'll pursue an idea from Daniel Gilbert (2006; Gilbert et al. 1998) and Eric Mandelbaum (2019) that aspects of unconscious cognition constitute a "psychological immune system." The psychological immune system contains both (ir)rational and arational aspects, supporting deep rationalism while situating it in a broader conception of cognition and its functions.

First, I'll argue in section 2 that a core function for the psychological immune system—unconscious rationalization to reduce cognitive dissonance—should be thought of as involving logical inference, and that this falsifies unconscious associationism and views of inference defended by internalists like Boghossian. I'll then propose in section 3 that the psychological immune system should be expanded to include more than just cognitive dissonance. The psychological immune system is not one thing, but instead constitutes a suite of distinct cognitive mechanisms that fulfil a general function of maintaining stable motivation in the face of widespread awfulness and death. Some of these mechanisms involve unconscious rationalization and others involve arational mechanisms designed to minimize negative affect. The resulting picture integrates deep rationalism with nonrational mechanisms of attitude change by appeal to a shared function of keeping us motivated.

2. Unconscious inference: an existence proof

Boghossian sorts automatic transitions in thought into the conscious cases (which may nonetheless be quick and automatic; Shea & Frith 2016), and the unconscious cases which are "made by [one's] subpersonal cognitive mechanisms" and are therefore "just programmed and not under the thinker's rational control" (2018, 61). Our key question in what follows is whether there are unconscious transitions between unconscious attitudes that do not seem to be either merely associative or "just programmed" operations of some subpersonal system, such as the visual system.

The term 'association' can refer to (*inter alia*) a kind of mental structure or a kind of transition between mental states (Mandelbaum 2016). For example, one might argue that a negative attitude toward broccoli is an association between the concept BROCCOLI and a negative valence, such that tokening either the concept or the valence increases the probability of tokening the other (Gawronski & Bodenhausen 2006). Or one might explain movements in thought, such as priming, by positing an associative transition from one concept like NURSE to an associatively related concept like DOCTOR (Meyer & Schvaneveldt 1971; Anderson 1983; cf. Dacey 2019).

Associated states ought to "rub off" on one another: if a subject strongly associates salt with pepper and an experimenter presents pepper together with an arbitrary stimulus over and over, the subject ought to form an association between salt and the arbitrary stimulus. This is most clearly visible in experiments using "evaluative conditioning" paradigms (Hofmann et al. 2010): pairing a negatively

valenced image, like a cockroach, with a neutral stimulus, like an unfamiliar Pokemon, causes subjects to associate the neutral stimulus with a negative valence (Olson & Fazio 2001). One should only be able change the valence in an associative structure from negative to positive by repeatedly introducing a positively valenced stimulus, a process known as counterconditioning. Since they lack any sort of logical form, both associative structures and associative transitions are resistant to counterevidence (Mandelbaum 2016).

A genuinely *inferential* process, understood in contrast with association, ought to involve a non-associative transition between non-associative states. Paradigmatically, inference involves moving from one *belief* to another (Harman 1986; Boghossian 2014), i.e., moving between mental states that are non-associative, truth-evaluable, and involve some sort of predicate-argument structure. While one might associate BANANA with YELLOW, the belief that bananas are yellow takes the additional step of *predicating* yellowness of bananas. The kind of movement also matters. One could imagine a purely associative transition between beliefs, while an inferential transition must show sensitivity to evidential relations between beliefs (Boghossian 2014; Mandelbaum 2016; Mody & Carey 2016).[1] Useful paradigm cases include transitions that instantiate a logically valid schema such as *modus ponens*, as in Boghossian's example of inferentially moving from IT'S RAINING and IF IT'S RAINING THEN THE STREETS ARE WET to THE STREETS ARE WET.[2] Part of the reason this transition is an inference rather than an associative transition or instance of a hard-wired rule is that it is sensitive to changes in the strength of its premises; if we come to believe that it is not the case that it's raining, the logical contradiction between this new belief and one of the premises shortcuts the inference altogether. To find a case of unconscious inference, then, we ought to find evidence for transition between beliefs that violates associative principles and is sensitive to the evidence available in a particular context, including evidence against one or more of its premises.

The case study used in this section is rationalization: specifically, cognitive dissonance reduction (Festinger 1957; Aronson 1992; Cooper 2005). According to the classic idea of cognitive dissonance, "nonfitting relations among cognitions" (Festinger 1957, 3) generate a negatively valenced affective state that motivates thinkers to reduce it by shifting around their attitudes. This negative motivational state is a feeling of dissonance.[3] Festinger regarded dissonance as a drive similar to hunger,

[1] As far as I know, there is no clear experimental evidence for associative transitions between entire propositional thoughts rather than individual concepts (like DOCTOR–NURSE). Intuitive examples include Boghossian's (2014) imaginary depressive who moves from I'M HAVING SO MUCH FUN to BUT THERE IS SO MUCH SUFFERING IN THE WORLD and Mandelbaum's (2016) imaginary Brit who moves from IT'S 10PM to I SHOULD GO TO THE PUB.

[2] There is no commitment here to the idea that logical inference in humans is always or even usually classically valid—there may instead be a proprietary mental logic (Braine & O'Brien 1998).

[3] Festinger used "dissonance" to refer to the presence of a contradiction in thought, whereas I will use it to refer to the affective state induced by contradiction.

with the shift in attitudes in order to assuage dissonance functioning similarly to the search for food to assuage hunger.

The notion of dissonance is best understood through experimental examples. Festinger and Carlsmith (1959) asked subjects to perform a boring, pointless task, and afterwards paid them either $1 or $20 to tell another person that the task was fun. Finally, they were asked what they honestly thought about the task. Paradoxically, the subjects who were paid *less* money reported a *more positive* attitude toward the task.

From an associationist perspective, this effect is hard to explain (Mandelbaum 2016). More money means more positive affect, which should associatively rub off on the task; why then do better-paid subjects think the task is worse? The dissonance-based answer is that the subjects realize that $1 is not really enough money to justify lying to another person. The fact that the task was boring conflicts with their undermotivated action of saying that the task was fun; this conflict generates a feeling of dissonance. The unpleasant feeling can be alleviated simply by shifting their attitudes—if I thought the task was actually fun, then my saying it was fun doesn't generate any conflict between my attitude and my behavior. So, the explanation goes, I form the belief that the task was fun after all. Some such non-associative story is required to explain how *less* reward can create a *more* positive attitude, a reversal that violates the most basic principles of associative learning.

Effects like this are domain general. The more unpleasant the initiation ritual, the more subjects like the (objectively quite boring) discussion group they've joined (Aronson & Mills 1959; Ma et al. 2014). Choosing between two equivalent items causes you to devalue the item you didn't choose (Brehm 1956; Lieberman et al. 2001). Voluntarily writing an essay against your own opinion on some topic causes you to change your opinion (Brehm & Cohen 1962; Heitland & Bohner 2010). The contradiction between the standing belief that p and the knowledge that you just acted as if not-p induces dissonance, which you may alleviate by shifting your belief toward not-p.

Crucially, these effects are typically mediated by beliefs about ourselves. Aronson writes that "at the very heart of dissonance theory, where it makes its clearest and neatest prediction, we are not dealing with any two cognitions; rather, we are usually dealing with the self-concept and cognitions about some behavior" (1969, 27). The "self-concept" consists of beliefs, principally (in healthy subjects) the beliefs (a) that we are competent (e.g., smart), (b) that we are good (e.g., moral), and (c) that we are stable (Aronson 1992). The fact that beliefs about the self are involved explains why the above effects are only observed when subjects take themselves to have freely chosen to engage in the relevant behavior. If they instead feel that they've been compelled by experimenters to (e.g.) write an essay opposing their own opinion or participate in a boring task, they can alleviate dissonance by blaming the experimenter; it's not that I behaved irrationally, it's that I had no choice.

Both the generation and the reduction of cognitive dissonance are (at least often) due to unconscious transitions between beliefs that are sensitive to logical form. Consider the generation of dissonance first, which often takes the following form:

(1) This task is not worthwhile.

(2) If this task is not worthwhile, then I am incompetent for freely choosing to engage in it.

(3) I am incompetent for freely choosing to engage in this task.

The conclusion, (3), contradicts the belief that I am a smart, competent person (a standing element of the self-concept). This contradiction generates the feeling of dissonance. The process of reducing dissonance is yet another inference that takes the following form:

(4) I am smart.

(5) Smart people only freely choose to do things that are worthwhile.

(6) I freely chose to engage in this task.

(7) The task must be worthwhile.

Believing (7) eliminates the contradiction, and with it the unpleasant feeling of dissonance.[4]

The case for a non-associative theory of cognitive dissonance is strong, and indeed the discovery of dissonance effects played an important but often neglected role in the downfall of behaviorism and its reinforcement-based approach to human behavior (Aronson 1997). However, some non-associative transitions might mimic the logical form of an inference without being literally inferential. To borrow Boghossian's example again, we might model the visual system's computation of depth via stereopsis as having a form like: *If the disparity between two retinal images for some point in the environment is X, then the point has depth value Y; These retinal points have disparity X; Therefore, the distal point has depth value Y*. One reason for doubting that the relevant computation is actually an inferential transition of this kind is that it is not sensitive to logical form. For example, acquiring counterevidence that undermines one of the "premises"—such as being told that you're looking at a random-dot stereogram, and thus that binocular disparity fails to indicate depth—has no effect on stereoscopic depth perception. Barring some independent reason to regard the transition as inferential, we ought to regard this as a subpersonal, "just programmed" (Boghossian 2018, 61) transition from a representation of binocular disparity to a representation of depth.[5]

[4] The fact that our minds contain mechanisms that hunt for and eliminate contradiction and the fact that humans tend to harbor contradictions jointly suggest that belief storage is *fragmented* (Quilty-Dunn & Mandelbaum 2018).

[5] Exactly what kind of evidence one would use to posit an inferential transition in subpersonal perceptual processing is an underexplored question (cf. Orlandi 2014; Jenkin 2020). For example, a transition within a module might fail to respond to counterevidence not because it lacks inferential form, but simply because it's encapsulated from the relevant

We don't find the same sort of hard-wired, evidence-insensitive character in the case of rationalization. As I modeled the transitions above, both the arousal of dissonance and the inferential means of reducing it depend on the belief that I am smart. In that case we can predict that increasing the salience of evidence against that premise will minimize dissonance effects. This prediction turns out to be true. Administering bogus personality tests and giving subjects the results that they are irrational and immoral eliminates dissonance effects (Glass 1964; see also Stone & Cooper 2003). Thus the transitions involved in rationalization do not merely mimic the logical structure of an inference, as a subpersonal perceptual process might. Instead, they rely on beliefs (e.g., about the self) as premises and are sensitive to evidence against those premises.

Dissonance reduction thus seems to involve unconscious inferences in order to preserve cherished beliefs about the self. Some cognitive scientists who take a broadly rationalistic (e.g., Bayesian) approach to cognition accept that these effects are explained by unconscious inferences, but have argued that the relevant transitions are rational, non-self-serving inferences about one's own attitudes. For example, one may reason as follows:

(8) I freely chose to engage in this task.

(9) If I freely chose to engage in this task, it must be because I think the task was worthwhile.

(10) I think the task is worthwhile.

This "self-perception theory" was originally put forth by Bem (1967; Bem & McConnell 1970; see also Cushman 2020). As Cushman puts it, "Rationalization is rational" (2020, 1).

Rationalistic self-perception approaches have trouble explaining the relevance of self-esteem to rationalization (note that the belief that I'm smart plays no role in the transition just outlined). But even more direct evidence tells in favor of an irrationalistic dissonance-based approach. Unlike the self-perception theory, the dissonance-based explanation critically posits a negatively valenced feeling of dissonance. This prediction also turns out to be true (Elliot & Devine 1994; Kitayama et al. 2013). Subjects show increased galvanic skin response when dissonance is posited to occur (Croyle & Cooper 1983; Elkin & Leippe 1986). Subjects with the option to drink alcohol fail to show attitude change in dissonance paradigms, likely because the alcohol mitigates the negative affect of dissonance and thereby makes attitude change redundant (Steele et al. 1981). Moreover, affective states can be modulated by how subjects conceptualize them (Schachter & Singer 1962). Prompting subjects to reconceptualize dissonance should therefore eliminate attitude changes characteristic of dissonance reduction, since the negative affect will no longer be experienced as dissonance. One test of this prediction involves simply administering a placebo and telling subjects it will make them

counterevidence. Since present purposes don't require taking a stand on this difficult methodological issue, I set it aside here.

uncomfortable. When dissonance is induced, subjects misattribute the negative affect to the placebo and fail to draw any dissonance-reducing inferences (Zanna & Cooper 1974). The self-perception theory can't make sense of the presence of negative affect in dissonance effects, or how modulating affect impacts the inferences subjects draw.

One might allow that dissonance exists and that it motivates attitude change but deny that it involves belief. In inference, "you start out with some beliefs, and either end up adding a new belief, or losing some beliefs you already had, or modifying the credence with which you hold some belief, or changing the basis on which you hold some belief" (Boghossian 2018, 56). It is hard to deny that the *generation* of dissonance involves belief. Dissonance arises when we notice that something we have done is in some way bad or irrational, and this causes a feeling of dissonance because it threatens our beliefs that we are neither bad nor irrational. The positive beliefs about the self that underlie healthy self-esteem in effect make certain conclusions *psychologically unacceptable*. When incoming evidence contradicts one of these core beliefs, the "psychological immune system" (Gilbert 2006; Mandelbaum 2019) responds with a pain-like state that motivates us to neutralize the invading inference. Thus the function of inferential belief-updating in cases of rationalization is at odds with rationality: instead of apportioning our beliefs to the evidence, we draw inferences and shift beliefs around in order to make ourselves feel better. Reasoning that demonstrates this self-serving character is a paradigmatic form of irrationality.

If we don't understand the generation of dissonance as arising from contradictions between beliefs, it's unclear why convincing someone that their core beliefs are wrong (e.g., that they are bad/irrational) should reduce dissonance (Glass 1964). Moreover, depressed subjects with low self-esteem are less likely to be subject to an "illusion of control"—when asked to press a button and determine to what extent their action causes a light to turn on, healthy subjects overestimate their degree of control while depressed subjects can be more accurate (Alloy & Abramson 1979; cf. Yon et al. 2020). This "depressive realism" is often overstated, since depression can also warp cognition in negative ways that engender false beliefs (Beck 2008; Carson et al. 2010). But the effect of depression minimizing self-aggrandizing cognitive illusions is quite real across many studies (Moore & Fresco 2012). Since long-standing depression or short-term blows to self-esteem weaken core beliefs like I AM COMPETENT or I AM MORALLY GOOD, they minimize dissonance by minimizing contradiction between incoming evidence and these core beliefs.

Still, while one might accept that the generation of dissonance is due to contradictions among beliefs, one might reject the description of the *reduction* of dissonance as a change in belief. Perhaps what I described above as subjects believing a task is worthwhile is really a nondoxastic state of liking/disliking. Many dissonance studies do involve positive or negative attitudes, which one may hesitate to call beliefs. Perhaps dissonance reduction is really about shifting attitudes rather than inferential belief change. I'll now describe some evidence that dissonance reduction involves changes

in descriptive belief. These examples highlight the intuitively irrational character of dissonance reduction.

Many examples of changes in descriptive belief to reduce dissonance are relatively benign—e.g., subjects who freely choose to skateboard up a hill on their knees will systematically underestimate the slope of the hill, thereby convincing themselves that it was not such an irrational decision (Balcetis & Dunning 2007).[6]

A more systemic example is the denial of racism by members of privileged racial groups. In general, African Americans are more knowledgeable than white Americans about the history of racism in the United States, and knowledge of this history is positively correlated with a tendency to judge events to be caused by racism (Nelson et al. 2013). Interestingly, a stronger sense of racial identity among white participants predicts greater denial of systemic racism as well as minimizes the effect that new knowledge about racism has on increasing perceptions of racism (Bonam et al. 2019). The fact that the effect is driven by the sense of racial identity suggests that the effect is ultimately driven by dissonance; evidence of racism only needs to be avoided or discounted if it causes dissonance by threatening the sense of self. Indeed, for white Americans with a high sense of racial identity, merely *thinking* about white privilege *increases* the strength of racist beliefs (Branscombe et al. 2007). Among white subjects who believe that the U.S. racial hierarchy reflects ethnic-group differences in ability and work ethic, being reminded of increased status of African Americans causes an increased belief in anti-white discrimination; this belief update improves self-esteem (Wilkins et al. 2017), and fails to occur in subjects who are given an opportunity to affirm their self-esteem in unrelated ways (Wilkins & Kaiser 2014).

The question of whether white privilege exists is a purely descriptive matter, and white Americans often answer negatively to preserve self-esteem, particularly when (e.g.) they identify as people who have achieved success through merit (Knowles & Lowery 2012; see also Knowles et al. 2014). The beliefs being formed to reduce dissonance concern factual, historical questions about the absence of racism and privilege, and their formation amounts to an active preservation of ignorance to avoid psychological distress. As Charles Mills puts it in a discussion of "white ignorance":

> *Ignorance is usually thought of as the passive obverse to knowledge, the darkness retreating before the spread of Enlightenment.*
> *But…*
> *Imagine an ignorance that resists.*
> *Imagine an ignorance that fights back.*

[6] Balcetis and Dunning (2007) take this effect to be mediated by cognitive penetration of visual perception. Firestone & Scholl (2016) mount a compelling critique against this interpretation of results like Balcetis and Dunning's; it's more likely that the effect is on belief alone.

(Mills 2007, 11).

Another case of dissonance reduction through changes in descriptive beliefs concerns meat. In consumer research journals, researchers have puzzled for the past decade over the "Meat Paradox": "people simultaneously dislike hurting animals and like eating meat" (Loughnan et al. 2010, 156). The fact that I eat meat while knowing it to be morally wrong creates dissonance by threatening my core belief that I am morally good. Subjects who just ate beef jerky are more likely than subjects who just ate cashews to deny that cows are capable of various forms of cognition (Loughnan et al. 2010). Merely reading a story about an individual in which it is briefly mentioned that he is vegetarian (as opposed to gluten-free, for controls) is enough to weaken meat-eating subjects' beliefs that animals experience emotions (Rothgerber 2014). Participants asked to offer justification for meat eating often rely on the "Four Ns": eating meat is (1) necessary (e.g., for protein), (2) natural (i.e., humans are meant to do it), (3) normal (i.e., humans generally do it), and (4) nice (i.e., meat tastes good) (Piazza et al. 2015). Subjects who strongly endorse the "Four Ns" (compared to those who don't) enjoy the benefits of cognitive dissonance reduction, experiencing "less guilt about their dietary practices" (Piazza et al. 2015, 123).

Rationalization about meat-eating is filtered through the self-concept. Men are more likely to incorporate meat-eating into their identity than women (Rothgerber 2013) due to associations between meat-eating and masculinity (Ruby & Heine 2011; Rozin et al. 2012). Men are correspondingly more apt to endorse the "Four Ns" than women (Fagerli & Wandel 1999; Piazza et al. 2015), likely due to their increased need to reduce dissonance. Upon viewing a video about how lambs are killed to make meat, men were more likely than women to react by not only offering justifications for meat-eating, but in fact by *increasing their commitment* to eating meat (Dowsett et al. 2018). Women, however, were less likely to reduce dissonance effectively in response to the video, showing a persistence of negative affect (Dowsett et al. 2018).[7]

In all these cases, the reduction of dissonance involves a change in descriptive belief. This sort of change in belief is not a form of rationally good inference: it is instead a form of *rationalization* (D'Cruz 2015; Schwitzgebel & Ellis 2017). However, rationalization is still inference. A theory of inference cannot restrict itself to good cases—bad inferences are inferences just the same. In these cases, we change the strength of our beliefs or adopt new beliefs in response to the strength of the evidence. The change in belief can be modulated by modulating premises (e.g., increasing or decreasing the strength of beliefs about ourselves).

[7] Dowsett et al. report that women are more likely to reduce dissonance by mentally dissociating meat from animals and underreporting their own meat consumption. The former strategy is not possible given the video, so it is harder to reduce dissonance; thus the dissonance simply persists, causing reports of higher negative affect.

Mere associative processing cannot explain how people change their beliefs to reduce cognitive dissonance. The effects are richer and more directly sensitive to evidence than that. Being reminded that eating meat is harmful should cause a negative valence to rub off onto the concept MEAT, but instead we respond defensively by rationalizing that it must not be that harmful after all; or that it is harmful but animals don't really have complex mental states and thus don't matter so much; or that it is harmful and animals have mental states but eating meat is still necessary for proper nutrition. Crucially, which strategy we adopt is sensitive to our evidence at the time. For instance, while subjects reminded of a committed vegetarian respond by decreasing their belief in animal minds, subjects reminded of a vegetarian who regularly lapses and eats meat instead respond by increasing their belief that it is too difficult to avoid eating meat (Rothgerber 2014). Thus rationalization exhibits the flexible sensitivity to salient evidence that is characteristic of domain-general inference.

Boghossian understands unconscious inferences on the model of subpersonal computations in the visual system. But that model is ill-suited to account for the generation or reduction of dissonance. Here the mental states involved are beliefs, and the beliefs are acquired, rejected, or modulated in strength in a manner that is proportional to the strength and content of incoming evidence (albeit in an epistemically distressing direction). Neither the states nor the processes involved are usefully described as "subpersonal" (except in the uninteresting sense that they are unconscious). These are the beliefs of an individual, and beliefs that play an important role in making rational sense of how they behave. And the sense-making sensitivity of beliefs to evidence is a paradigm case of a personal-level mental operation.

The case against regarding unconscious transitions as genuine inferences, for Boghossian, is grounded in questions about responsibility. "[F]or it to make sense to hold you responsible for your inferences, inferring has to be something you *do*, and not just something that happens to you" (Boghossian 2018, 60). And we don't hold people responsible for "sub-personal cognitive mechanisms" which are "just programmed and not under the thinker's rational control" (Boghossian 2018, 61).

One question is: What properties of an unconscious belief change could render a thinker responsible for it? That's a theoretical question about the notion of responsibility that may rely, for example, on a notion of epistemic basing on reasons (Jenkin 2020). That theoretical question is not taken up here. But another, more immediately tractable question concerns particular cases: Do we intuitively judge people to be responsible for some unconscious belief changes? Answering this question can provide counterexamples to the thesis that inferences are constitutively conscious actions even in the absence of alternative conceptions of responsibility and rational evaluability.

Consider the case in which a white American responds to evidence of racism with *decreased* belief in white privilege, or in which a meat-eater responds to evidence of the possibility of a vegetarian lifestyle with decreased belief in the capacity of animals to feel suffering, or in which a person convinces

themselves that an object they built is more valuable simply because they invested so much time in making it (the so-called "IKEA effect"—Norton et al. 2012). I suggest that intuition does not ally these processes with the stereoscopic computation of depth. Instead, we are inclined to judge these subjects as engaging in epistemically degenerate rationalizations, and thereby hold them epistemically responsible. A white person responding to salient evidence of white privilege by decreasing their belief in racism in order to avoid unpleasant feelings is about as clear an example of irrationality as one could hope to find.

Furthermore, the intuitive irrationality (and thus rational evaluability) of rationalization fits with its underlying cognitive mechanisms, which exhibit just the sort of flexible evidence responsiveness that cannot be found in merely associative or hard-wired subpersonal processes. We can also point toward a distinct, non-rational function being fulfilled by rationalization processes: neutralizing threats to self-esteem by reducing a negatively valenced drive state. Rationalization thus involves shifting beliefs, for self-serving reasons, away from the direction pointed to by the incoming evidence. These rationalizations often interact with, and contribute to, prevailing ideologies in ways that we justly hold each other responsible for. The notion of rationally evaluable inference ought to be invoked to make sense of this cognitive activity, even if its unconsciousness makes it difficult to square with more general theses about the conditions of responsibility.

3. Boosting the psychological immune system

3.1—Awfulness and death. The foregoing has pointed away from unconscious associationism toward deep rationalism. It also raises questions about the proper function of our attitudes and the means by which we update them. For some philosophers, belief "aims at the truth" (Velleman 2000, 244; cf. Bortolotti 2020); the ultimate point of reasoning is "to get things right" (McHugh & Way 2018, 178). But the updating mechanisms underlying rationalization don't seem to have this aim. Instead, they seem to involve what Mandelbaum calls a *psychological immune system*: "the beliefs one changes (or keeps) are due to what feels easiest to do while keeping one's self-image intact" (2019, 153). But why should our minds have this immune system built in?

For Gilbert (2006), the point of the psychological immune system is to keep us happy, even when life takes a downturn. The need for this kind of system plausibly arises from a sort of design problem for human minds. We are rational creatures capable of gaining general knowledge about the world and our place in it, and we are also creatures that require stable motivation to keep moving. However, there are two general features of life on Earth that threaten stable motivation, which I will unceremoniously label *awfulness* and *death*. The first, awfulness, consists of the ambient sources of harm and desire-frustration that populate our environments. Awfulness is responsible for a range of negative experiences from minor pains to the frustration of strong desires to life-ruining misfortune.

Awfulness can be *internal* (negativity due to the self, e.g., immoral behavior) or *external* (negativity due to the world, e.g., harmful circumstances). A rational creature capable of catching onto pervasive awfulness is in danger of being overcome with negative affect, resulting in the maladaptive "low motivation for engaging with the outside world" characteristic of depressive disorders (*DSM-5*, 194; see also Wang et al. 2006; Sherratt & MacLeod 2013).

The philosophical tradition of pessimism takes awfulness to be the dominant feature of life: "the quality of even the best lives is very bad" (Benatar 2006). While even pessimists grant that life is not universally awful, it is not obvious that the joy-giving aspects of the world outweigh the awfulness. Consider Schopenhauer's stark juxtaposition of the intense suffering of the animal being eaten and the comparatively minor pleasure enjoyed by the animal doing the eating (1851, 292). We face a real predicament of why we ought to keep going when we have so little control over what happens in the world, our hopes are so often dashed, and experiences often range from the dreadful to the merely boring (Benatar 2006; 2017).

We need not endorse pessimism to see this point—the sheer volume of awfulness is enough to pose a motivational conundrum even if pessimism overstates its prevalence. And even if awfulness fails to predominate today (see Pinker 2011 for a rosy view of recent history), it surely did at various points in the past. Scarcity of resources was likely a constant in our evolutionary history. Some geneticists and paleoanthropologists hypothesize that the human population may have at one point dwindled down to as few as 1,000 members (Hawks et al. 2000; Li & Durbin 2011). Aspects of a psychological immune system may have evolved to cope with our brutal past. Another feature of life on Earth, which unlike awfulness has been perfectly general at every point in history, is the inevitability of death (on which more below).

How could you design a rational mind in a way that enabled it to maintain stable motivation in the face of ubiquitous awfulness and death? A psychological immune system that allows for unconscious rationalization to preserve a stable self-concept provides part of the answer. Allowing constant frustration of desires to make us believe we are insignificant and lack control over our futures threatens to make us devalue our own desires and become pathologically unmotivated. If we instead believe ourselves to be valuable, competent, and in control and experience dissonance when we receive evidence to the contrary, then we are driven to preserve our cherished beliefs and, if necessary, to irrationally increase our beliefs in our own value, competence, and control in the face of disconfirmatory evidence.

Understood in the context of the psychological immune system, some cases of depression can be thought of as a kind of immunodeficiency. As noted above, depression often involves a systematic warping of belief (Beck 2008). But insofar as it undermines the self-concept and thereby prevents the

generation of dissonance, it may reduce the likelihood of unconscious rationalization. This unbiasing aspect of depressive realism was anticipated by Freud in his description of the "melancholic" patient:

> [I]t is merely that he has a keener eye for the truth than other people who are not melancholic. When in his heightened self-criticism he describes himself as petty, egoistic, dishonest, lacking in independence, one whose sole aim has been to hide the weaknesses of his own nature, it may be, so far as we know, that he has come pretty near to understanding himself; we only wonder why a man has to be ill before he can be accessible to a truth of this kind.
>
> (Freud 1917, 246)

For Mandelbaum, the self-concept and the dissonance it generates constitute the core of the psychological immune system. But the motivational problem created by awfulness and death are not limited to the self. You might believe that you are a good person whose desires are worth pursuing, but the awfulness of the world around you persists. Even if you're fantastic, why bother continuing in the face of a world that consistently delivers suffering and boredom? *External* awfulness (i.e., awfulness attributed to the environment rather than to the self) mounts a more or less constant threat to motivation that cannot obviously be defeated by a strong self-concept alone.

Another objective problem that cannot obviously be quelled through dissonance is death. It is compatible with your being a good, rational, competent person that you will inevitably die. The knowledge of our own mortality represents perhaps the greatest design problem that faces us as rational creatures. Solomon, Greenberg, and Pyszczynski identify this problem in terms of the intense anxiety that accompanies the recognition of impending doom, which they call "terror":

> Terror is the natural and generally adaptive response to the imminent threat of death. All mammals, including humans, experience terror. When an impala sees a lion about to pounce on her, the amygdala in her brain passes signals to her limbic system, triggering a fight, flight, or freezing response.
>
> (Solomon et al. 2015, 7)

The design problem for rational, reflective creatures like us is our *stimulus-independent* ability to recognize that we will die. A human being, unlike a cat, can sit comfortably on a couch with no danger in sight and still be troubled by thoughts of death. A rational creature with no defense mechanisms against such thoughts is in danger of living in permanent terror, creating a massive motivational problem. And again, dissonance reduction alone seems ill-equipped to cope with this threat.

I propose (*pace* Mandelbaum) that we expand the concept of the psychological immune system beyond the tendency to generate dissonance in response to incoming contradictions with the self-concept. The psychological immune system is not a single mechanism serving a single function. Instead, it constitutes a diverse array of distinct mental mechanisms keyed toward the general

maintenance of stable motivational structures in the face of awfulness and death. One core component is indeed dissonance and the preservation of the self-concept it affords. But other components outstrip the self-concept, concerning the evaluation of objective circumstances and death-related cognition.

3.2—Death. Consider death-related cognition first. Solomon et al. (2015) posit a mechanism of *terror management* that minimizes the experience of death anxiety. Thoughts of death are like a contagion, threatening to infect our minds with paralyzing anxiety. Terror management involves quarantining these harmful thoughts by pushing them outside of consciousness and making them harder to access. A primary cognitive mechanism of managing death anxiety—indeed another example of unconscious rationalization—is to negate the inevitability of death by bolstering the belief in immortality. Immortality can be literal, in which case people may cling to religious beliefs in the persistence of the self after bodily death, or symbolic, in that we can metaphorically "live on" through our participation in and contribution to a meaningful community that outlives us (Pyszczynski et al. 2015). People therefore respond to thoughts of death through *worldview defense*: we push thoughts of death out of consciousness by reaffirming the meaningfulness of our lives in our communities (religious, nationalistic, etc.).

Results supporting this hypothesis include asking people whether they prefer products made by their own country. German participants show increased preference for all things German if interviewed in front of a cemetery rather than in front of a department store (Jonas et al. 2005). Israeli children reminded of death are more likely to prefer to play with other Israeli children over Russian children (Florian & Mikulincer 1998). After subliminally seeing death-related words (controls saw pain-related words), American participants presented with an essay critiquing the U.S. and an essay praising it find the arguments in the former to be weaker and in the latter to be stronger (Arndt et al. 1997a).

Modulation of meaning-giving sets of beliefs (or "worldviews") correspondingly modulates the psychological accessibility of death-related thoughts. The accessibility of death-related thoughts is often measured through word-stem completion tasks. For example, suppose you were asked to fill in the missing letters in "COFF_ _". You might complete the stem to yield the benign word "COFFEE", or to yield the death-related word "COFFIN". Subjects who are reminded of death and then given an opportunity to strengthen the beliefs underlying their worldview (i.e., by judging pro-American arguments to be valid) show reduced accessibility of death-related concepts; they are significantly less likely to provide "COFFIN" (Arndt et al. 1997b).

Similar "mortality salience" effects are found in the opposite direction: providing evidence against a person's worldview increases the accessibility of death-related concepts. Religious participants who deny evolution and read evidence in favor of evolution show greater accessibility of death-related

concepts in the stem completion task (Schimel et al. 2007). Canadian subjects who read an essay by an American author belittling hockey and other sources of Canadian pride show the same result (*ibid.*).[8]

As aforementioned, it's hard to see how terror management could reduce to dissonance reduction. Evidence that we will die does not contradict our beliefs that we are good, competent, and consistent. Moreover, while dissonance reduction can involve shifting around our attitudes, terror management also accomplishes the separate task of modulating the accessibility of death-related concepts. Dissonance may inhibit retrieval of beliefs (e.g., about the boringness of the task I volunteered to perform), but has not been shown to inhibit retrieval of concepts independent of their role as constituents in unpleasant beliefs—terror management, on the other hand, inhibits retrieval of death-related concepts independently of the attitudes they figure in. Some terror management effects

[8] Concerns about terror management have arisen in the current "replication crisis" in cognitive science. Sætrevik & Sjåstad (unpublished) uploaded a pre-registered replication attempt on PsyArXiv in May 2019 showing a failure to replicate mortality salience effects in Norwegian subjects, later supplemented with a failure to replicate effects on American subjects in an online study. The introduction of novel experimental contexts may introduce novel confounds, e.g., the probable difference in nationalistic sentiment in Norway vs. the United States. The authors also used pro-democratic essays for Norwegian participants, but those essays argued against increased control in response to the threat of terrorism; it is perhaps unsurprising that even robustly pro-democratic Norwegian participants may, when mortality is salient, hesitate to agree with the following (quoted from the authors' English translation of their materials): "increased terror preparedness and increased control is not the solution for the future of Norway!" The study of Americans was performed online, and thus may not constitute ideal experimental conditions for mortality salience (Chopik & Edelstein [2014] did find evidence of highly significant mortality salience effects in an online study with large sample sizes, though unlike Sætrevik and Sjåstad their study was specifically focused on internet-based cues to mortality salience).

More apparent replication failures emerged in December 2019 from Many Labs 4 (Klein et al. unpublished), wherein 21 labs ran experiments on over 2,000 subjects and failed to replicate a mortality salience effect on worldview defense (Greenberg et al. 1994), despite earlier successful replications (e.g., Arndt et al. 1997a) and direct guidance from the original authors in many of the replication attempts.

Chatard, Hirschberger, and Pyszczynski (unpublished) argue that Klein et al. erroneously included experiments with sample sizes below preregistered inclusion criteria. In their re-analysis, Chatard et al. found that, when analysis is restricted to original-author-advised experiments that met preregistered sample sizes and Klein et al.'s exclusion criteria (e.g., only white American subjects), the effect was successfully replicated after all. Chatard et al. suggest that the apparent replication failure was "likely driven by a few small, heterogeneous, and imprecise studies that should not have been included in the meta-analysis if the authors had conducted the studies as planned." The replicated effect size was very small. One possible explanation for this is that the participants were college students tested in Fall 2016 and Spring 2017, during and right after Trump's election and inauguration. It's possible that a young, liberal population was alienated from U.S. political culture during that time such that strengthening pro-U.S. attitudes was a less viable means of terror management, causing decreased effect size.

These papers remain unpublished and proper methodology is a subject of ongoing debate amongst the authors—and in discussions of replicability more broadly. There have been over 400 experiments published showing mortality salience effects (Burke et al. 2010; 2013). These partial replication failures don't justify discounting the legitimacy of those effects, particularly given Chatard et al.'s argument for successful replication. More replication attempts are needed that show sensitivity to the social/historical context of the population being tested and the beliefs they're most likely to cling to.

do resemble dissonance effects in that cherished beliefs are challenged by counterattitudinal information and provoke a cognitive response. But the function of re-organizing cognition to inhibit retrieval of a particular family of concepts that are semantically unrelated to the counterattitudinal information seems *prima facie* to implicate a distinct mechanism from dissonance.[9]

Finally, dissonance reduction is driven by affect; rationalization is aimed at alleviating the unpleasant feeling of dissonance, which shows up in galvanic skin response (Croyle & Cooper 1983). But reminding subjects of death "showed no hint of either elevated self-reports of fear or anxiety and no increase in autonomic or cardiovascular indicators of arousal" (Pyszczinsky et al. 2015).[10] This fact suggests that the rationalization underlying terror management is not about reducing present negative affect, but instead about reorganizing cognition to avoid the *potential* for negative affect (Greenberg et al. 2003). In that case, terror management and dissonance reduction are fundamentally distinct cognitive mechanisms.

However, both mechanisms function as aspects of the psychological immune system and can work in tandem. For instance, while subjects who choose to write a counterattitudinal essay show a subsequent across-the-board preference for information supporting the essay (a standard dissonance effect), subjects who have been reminded of death show a significant increase in that preference (Friedman & Arndt 2005; see also Jonas et al. 2003). Thus reaffirming the self-concept can serve not only to reduce dissonance but also to reaffirm meaningfulness in the face of death. This fits with Solomon et al.'s insistence that self-esteem is "constantly at work, prodding us on beneath the surface of awareness to maintain our protective shield against terror" (2015, 47).

The self thus seems to play a significant role in mitigating death anxiety. Nichols et al. (2018) tested a range of subjects, including American Christians, American nonreligious subjects, Indian Hindus, and Tibetan Buddhist monks, for attitudes toward death and the self. They found that the Buddhist monks had not only (a) the lowest degree of belief in an enduring self, but also, surprisingly to the authors, (b) the *highest* degree of fear of self-annihilation caused by death. The Buddhist monks also were the most egocentric in a hypothetical tradeoff between months of one's own life against months/years of another's life (i.e., when life-extending medicine is in short supply). Lowering the

[9] Some dissonance theorists cite "repression" as a strategy for reducing dissonance, which may seem to suggest a similar mechanism to terror management (e.g., Zanna & Aziza 1976). Repression in this context has always been understood at the level of whole attitudes, however, not individual concepts. It is an interesting question for future research whether dissonance can drive people to repress concepts as well as attitudes—a positive answer would suggest that terror management and dissonance reduction may share cognitive mechanisms after all. Thanks to an anonymous referee for pressing this point.

[10] The claim that death reminders never cause negative affect is surely false (Lambert et al. 2014). But the lack of consistent evidence for affect suggests that the mechanism of terror management cannot be tied constitutively to the reduction of negative affect the way that dissonance reduction can.

strength of belief in the self appears to weaken aspects of the psychological immune system geared toward alleviating death anxiety.

3.3—Awfulness again. I've argued so far that the psychological immune system involves two separate but interacting components: cognitive dissonance and terror management. Cognitive dissonance serves the function of maintaining a robust self-concept and terror management serves the function of avoiding the paralyzing fear of death. But these two processes alone won't suffice to mitigate the negative affect coming from negative experiences unrelated to the self. A pessimist like Schopenhauer or Benatar would insist that the world delivers considerably more negative than positive. In that case we would need some aspect of our immune system to counteract this influx of negative affect and maintain stable motivation. Even if the negative and positive are perfectly matched in our world, we would still benefit motivationally from a positive skew; instead of viewing the world as coldly neutral, we could have some source of hope that things in general will be at least somewhat positive.

Another element of the psychological immune system may therefore be a *domain-general positive bias*. This need not be a single process but could instead be a general tendency of many cognitive processes to push evaluation toward the positive, all else equal. Some of these mechanisms might involve full-blown rationalization while others involve arational modulation of valence.

If there really is a domain-general positive bias, we might expect neutral items regularly to have a slightly positive valence. Concepts of ordinary objects and events are linked to "microvalences" (Lebrecht et al. 2012), and we should expect these to tilt toward positivity. One source of evidence for this comes from linguistic corpuses. Dodds et al. (2015) compiled a list of over 100,000 words from ten languages and acquired 5,000,000 valence ratings. They found a clear positive bias: the clear majority of both frequent and rare words across various languages are positively valenced (see Fig. 1). Looking at Warriner et al.'s (2013) English corpus, many perfectly boring words seem to have a slightly positive valence: e.g., driftwood (5.53), eraser (5.64), lamp (5.74), place (5.86), mild (5.9), initiate (6.1).

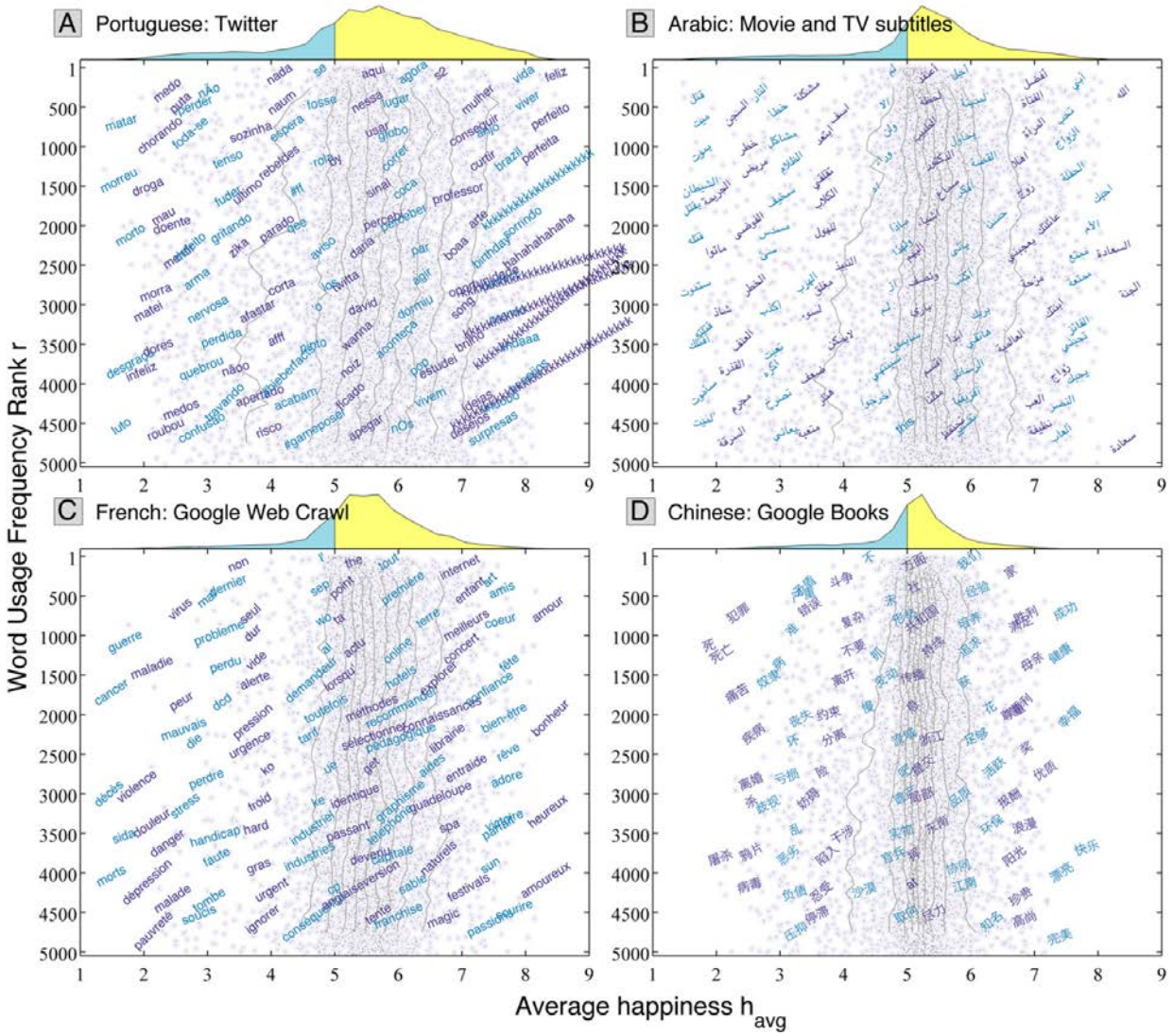

Figure 1—Valence assignments for linguistic corpuses skew positive for high- and low-frequency words alike (from Dodds et al. 2015).

Another example of domain-general positivity is the "mere exposure effect": merely being exposed to a stimulus makes subjects like that stimulus more, *ceteris paribus* (Zajonc 1968; Bornstein 1989). The mere exposure effect is fully domain general, applying to images and syntactic structures alike (Luka & Barsalou 2005).[11] The fact that the mind simply boosts valences toward the positive as

[11] An anonymous referee asks whether the mere exposure effect might explain why words tend to have positive valences. That hypothesis would seem to predict that the positive skew should strongly correlate with frequency, since higher frequency entails more exposure. But since the effect is "strongly independent of frequency" (Dodds et al. 2015, 2389), it seems there is a default bump in word valence that's not due to mere exposure.

objects are presented (*modulo* boredom and satiation [Montoya et al. 2017]) is a straightforward example of domain-general positive bias.[12]

There also exists a well-known *negativity bias*, captured by the slogan "bad is stronger than good" (Baumeister et al. 2001; see also Rozin & Royzman 2001; cf. Corns 2018). For example, an event-related potential observed in electroencephalogram for "oddball" stimuli is higher for a negative stimulus among positive stimuli than for a positive stimulus among negative stimuli (Ito et al. 1998). People show greater affective response to the possibility of losing $100 than gaining $100 (Kahneman & Tversky 1984; Zhao et al. 2020). Three-month-olds show a looking-time preference for helpful agents over harmful ones—a triangle-with-eyes struggling up a hill might be pushed up by a "helper" or downward by a "hinderer". But while three-month-olds show no preference for helpers over neutral agents, they show a marked preference for neutral agents over hinderers, suggesting that the effect is driven by a (possibly innate) social negativity bias (Hamlin et al. 2010).

Such negativity bias effects may seem to challenge the idea of a domain-general positive bias. But in fact, these biases complement each other—indeed they necessitate each other. First, their functions are not in competition. The positive bias primarily governs valence *assignment*, while the negativity bias governs valence *salience*: while valences skew positive, negative stimuli demand more processing resources and are more likely to be encoded and stored in memory (Ito & Cacioppo 2005). If we generally skew positive in our valence assignments, we're in danger of failing to recognize harmful stimuli. It would help in that case to have stimuli that cross a threshold of negativity to "pop out" and demand processing, thereby keeping the glow of positivity in check. Turning to memory, a positive skew in valence assignment makes it less pressing to store positive valences; instead, we can offload that burden onto our positive default valence assignments when re-encountering positive stimuli, and focus memory resources on encoding negative stimuli, as seen in the infant negativity bias (Hamlin et al. 2010).

In long-term memory, however, positivity tends to win out. While a bias toward negative stimuli for immediate salience and short-to-medium-term storage is functional, the general dominance of negativity in long-term memory could overload cognitive processes with negative affect, threatening motivation. Thus positive stimuli are often easier to recall than negative stimuli in long-term memory (Taylor 1991). Negative affect fades more than positive affect over 1-week intervals (Holmes 1970), a trend that continues over 3-month, 1-year, and 4.5-year intervals (Walker et al. 1997). In one recent

[12] A plausible explanation for the mere exposure effect is based in metacognition. Perceptual and cognitive processing of stimuli is accompanied by a metacognitive feeling of "fluency", a sense of easiness in processing (Alter & Oppenheimer 2009). As an object is encountered more and more, it feels more fluent to process it. Subjects may then attribute the positive valence of fluency as due to liking the object (Bornstein & D'Agostino 1994). In that case the mechanism underlying the effect may have a rationalistic structure. The data is messy, however, suggesting fluency may not be the (only) factor underlying the mere exposure effect (Montoya et al. 2017).

study, Red Sox and Yankee fans' memories were tested for details about game 7 of the American League Championship Series in 2003 (a devastating loss for the Red Sox and win for the Yankees) and 2004 (the reverse). Red Sox fans remembered the 2004 series better than 2003, while Yankee fans remembered 2003 better than 2004 (Breslin & Safer 2011).

The general picture that emerges here is one of a warm, fuzzy positive background in the mind, against which negative stimuli pop out, grab resources, and fade in the long-term. The concept of a slightly positive background may provide some motivation to be a "cognitive miser," i.e., to exhibit the human tendency to default to lazier, less demanding forms of cognition (Stanovich 2018).[13] While cognitive miserliness is often understood as a design feature that preserves resources, it might be actively motivated by the positive background hum of the mind—if the current mental terrain tends toward positive valence, why venture out into the unknown? And indeed, the phenomenon of "mind wandering" is a common source of momentary unhappiness (Killingsworth & Gilbert 2010), which seems to be due to cognitively stumbling over negatively valenced information while searching long-term memory (Poerio et al. 2013).

As the picture sketched here would predict, the balance between positivity and negativity biases is upset in depression: depressed subjects show greater accessibility for negatively valenced items in memory (Watkins et al. 1996; LeMoult & Gotlib 2019), memory deficits for positive items (Dillon 2015), and a reduction in fading of negative affect in memory (Walker et al. 2003). The positivity bias needs a corresponding negativity bias to avoid harm, and the negativity bias needs the positivity bias to avoid maladaptive dominance of negative affect.

3.4—Moving forward. The psychological immune system is not a single process. It instead reflects a general tendency of cognition to maintain stable motivation through a variegated array of cognitive mechanisms. Three kinds of psychological immune response include cognitive dissonance, terror management, and a domain-general positive bias, itself underwritten by multiple distinct mechanisms.[14] These aspects of cognition enable us to keep going in a world rife with awfulness and death.

---

[13] Thanks to an anonymous referee for suggesting this intriguing idea.

[14] Even dissonance may not be a single mental kind. The discussion above construed dissonance as based in conflict with the self-concept. But there may be a simpler kind of dissonance as well, in which mere contradiction is sufficient to generate dissonance independently of the self-concept. We cannot successfully act in the world if our beliefs are unstable and conflicted; thus there is some motivation to maintain consistency independently of the self (Harmon-Jones et al. 2009). There is some evidence for dissonance arising in nonhuman animals, for example, which may lack a robust self-concept (Harmon-Jones et al. 2017). In that case, dissonance may arise from two sources, one based in avoiding conflict and instability and the other based in protecting the self-concept.

Accepting this framework for thinking about human cognition opens up a research program. Other belief-related phenomena ought to be found that don't fit neatly into the categories described here but still play an immunodefensive role in that they possess four diagnostic properties: (1) they rely on representations of the world that adaptively enhance one's sense of playing a role within a meaningful community and/or increase the value attributed to oneself or one's circumstances; (2) encountering rational evidence that contradicts those representations triggers a motivation to respond cognitively; (3) the cognitive response is often biased toward preserving the pre-existing representations at the expense of truth and knowledge; and, I add tentatively, (4) the epistemic strength of these representations is often weakened (a) as self-esteem is weakened or (b) in cases of depression, anxiety, or other mental illnesses that involve negative appraisal of oneself and/or one's circumstances.

By way of illustration: another example of such a phenomenon is the "belief in a just world," i.e., the belief many people have that good things happen to good people and bad things happen to bad people (Lerner 1980; Hafer & Sutton 2016). This belief is *prima facie* false, but it paints a comforting picture of human outcomes in terms of moral desert rather than the alienating idea that good and bad outcomes often arise from uncontrollable independent factors. We want to believe that justice is not merely some Platonic ideal, but is in fact a causal force in determining human outcomes. This belief thereby blocks out an ambient source of awfulness, viz., *injustice*. As we would predict, people respond to evidence of the falsity of this belief not by decreasing its strength, but by irrationally adopting attitudes that are consistent with it even at the expense of inconsistency with the evidence. For example, even when subjects are told that punishment and reward will be doled out randomly, they nonetheless form positive appraisals of the character of those who receive rewards (Lerner 1965) and negative character-appraisals of those who receive random shocks (Lerner & Simmons 1966). People thus defend cherished beliefs in universal justice through irrational "victim-blaming."

The immunodefense of the belief in a just world also includes shifting ordinary descriptive beliefs away from truth. When told about a person who won the lottery, subjects who are told that the lottery-winner was a bad person later recalled his winnings as lower than subjects who were told that he was a good person (Callan et al. 2009). And, as an immunodefensive theory predicts, the belief in a just world is psychologically beneficial. In a survey of nearly 2,000 Europeans, subjects with stronger belief in a just world had greater perceptions of organizational justice in their workplace as a result, and accordingly showed greater workplace satisfaction a year later (Johnston et al. 2016). The effect is also modulated by self-esteem. Subjects who recall times they were lucky(/unlucky) respond with raised(/lowered) self-esteem; moreover, while healthy subjects typically regard their own unlucky outcomes as unfair, lowering their self-esteem causes them to regard their own unlucky outcomes as less unfair (Callan et al. 2014). Finally, as predicted, depression correlates with decreased belief in a just world (Ritter et al. 1990; Lipkus et al. 1996).

4. Conclusion

The popular idea that the unconscious mind is dumb and arational is false. The mind is not flat; unconscious processing has rich inferential structure. This fact has implications for how we understand inference (i.e., it requires neither consciousness nor "taking" premises to support conclusions) and human nature (i.e., it is deeply rationalistic). It also sheds light on the function of belief. Some philosophers argue that belief aims at truth (Velleman 2000) or knowledge (Williamson 2000). But insofar as the psychological immune system characterizes a function of belief, it looks as though belief often aims instead at preserving stable motivation even if truth and knowledge fail to be maximized (Mandelbaum 2019, 151; Bortolotti 2020). There must be a trade-off between these functions of belief, as it would clearly not be adaptive to fail to realize that a tiger is about to eat you simply because the recognition would cause negative affect (Aronson 1992, 30–31). But a view that construes belief as oriented toward truth (or knowledge) alone ignores the forms of irrationality that operate in the depths of human cognition—not as performance errors (such as slips of the tongue or arithmetical errors), but as systematic expressions of core cognitive competence.

Not all aspects of the psychological immune system are rational/irrational. The positive bias in valence assignment seems generally arational, as does the terror-management-based suppression of death-related concepts. Positing a psychological immune system allows us to defend a pluralistic form of deep rationalism on which unconscious inference is integrated with affect and arational cognitive mechanisms. This synthesis, which unites various rational and arational aspects of cognition by appeal to a common immunodefensive function, might be called "deep irrationalism".

The processes underlying the psychological immune system are beneficial in maintaining stable motivation and other respects (such as valuing things you've put effort into, like artifacts or friendships), but in other respects they are plainly irrational and even harmful. Thinking about the role that dissonance reduction plays in cognition about white privilege or that terror management plays in bolstering in-group preference is enough to make one wonder if the benefits are worth the costs. Indeed, the psychological immune system may help explain the formation, spread, and maintenance of ideology, which Adolph Reed, Jr. characterizes as "the mechanism that harmonizes the principles that you like to think you hold with what advances your material interest" (Reed 2014). Ideology may be best captured at the level of social groups (e.g., classes, races, and other loci of material interest) rather than individualistic psychology. But the psychological immune system helps explain how some ideological justifications take root in individual minds by relieving the psychological pressures of cognitive dissonance and death anxiety.

Crucial outstanding questions include how, and whether, to change these underlying cognitive tendencies. Possible avenues for mitigating the bad effects of the psychological immune system include taking care in what we allow ourselves to identify with (and thus what generates dissonance when

criticized) and which beliefs we rely on to derive meaning (and thus what gets strengthened when death is salient). Answers to these questions are, unfortunately, not easy to come by.[15]

References

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), 441–485.

Alter, A.L., & Oppenheimer, D.M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review* 13(3), 219–235.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th edition*. Arlington, VA: American Psychiatric Publishing. Cited as *DSM-5.*

Anderson, J.R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior* 22, 261–295.

Arndt, J., Greenberg, J., Pyszczynski, T., & Solomon, S. (1997a). Subliminal exposure to death-related stimuli increases defense of the cultural worldview. *Psychological Science* 8(5), 379–385.

Arndt, J., Greenberg, J., Solomon, S., Pyszczynski, T., & Simon, L. (1997b). Suppression, accessibility of death-related thoughts, and cultural worldview defense: Exploring the psychodynamics of terror management. *Attitudes and Social Cognition* 73(1), 5–18.

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

---. (1992). The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry* 3(4), 303–311.

---. (1997). Back to the future: Retrospective review of Festinger's *A Theory of Cognitive Dissonance*. *The American Journal of Psychology* 110(1), 127–137.

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology* 59, 177–181.

Balcetis, E., & Dunning, D. (2007). Cognitive dissonance and the perception of natural environments. *Psychological Science* 18(10), 917–921.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C., & Vohs, K.D. (2001). Bad is stronger than good. *Review of General Psychology* 5(4), 323–370.

Beck, A.T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry* 165(8), 969–977.

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

Bem, D.J., & McConnell, H.K. (1970). Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation attitudes. *Journal of Personality and Social Psychology* 14(1), 23–31.

Benatar, D. (2006). *Better Never to Have Been: The Harm of Coming into Existence*. Oxford: OUP.

[15] [acknowledgements]

---. (2017). *The Human Predicament: A Candid Guide to Life's Biggest Questions*. Oxford: OUP.
Boghossian, P. (2014). What is inference? *Philosophical Studies* 169, 1–18.
---. (2016). Reasoning and reflection: A reply to Kornblith. *Analysis* 76(1), 41–54.
---. (2018). Delimiting the boundaries of inference. *Philosophical Issues* 28, 55–69.
Bonam, C.M., Das, V.N., Coleman, B.R., & Salter, P. (2019). Ignoring history, denying racism: Mounting evidence for the Marley hypothesis and epistemologies of ignorance. *Social Psychology and Personality Science* 10(2), 257–265.
Bornstein, R.F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin* 106(2), 265–289.
Bornstein, R.F., & D'Agostino, P.R. (1994). The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Social Cognition* 12(2), 103–128.
Bortolotti, L. (2020). *The Epistemic Innocence of Irrational Beliefs*. Oxford: OUP.
Branscombe, N.R., Schmitt, M.T., & Schiffhauer, K. (2007). Racial attitudes in response to thoughts of white privilege. *European Journal of Social Psychology* 37, 203–215.
Brehm, J.W. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology* 52(3), 384–389.
Brehm, J.W., & Cohen, A.R. (1962). *Explorations in Cognitive Dissonance*. Hoboken, NJ: Wiley.
Breslin, C.W., & Safer, M.A. (2011). Effects of event valence on long-term memory for two baseball championship games. *Psychological Science* 22(11), 1408–1412.
Burke, B.L., Martens, A., & Faucher, E.H. (2010). Two decades of terror management theory: A meta-analysis of mortality salience research. *Personality and Social Psychology Review* 14(2), 155–195.
Burke, B.L., Kosloff, S., & Landau, M.J. (2013). Death goes to the polls: A meta-analysis of mortality salience effects on political attitudes. *Political Psychology* 34(2), 183–200.
Callan, M.J., Kay, A.C., Davidenko, N., & Ellard, J.H. (2009). The effects of justice motivation on memory for self- and other-relevant events. *Journal of Experimental Social Psychology* 45, 614–623.
Callan, M.J., Kay, A.C., & Dawtry, R.J. (2014). Making sense of misfortune: Deservingness, self-esteem, and patterns of self-defeat. *Journal of Personality and Social Psychology* 107(1), 142–162.
Carson, R.C., Hollon, S.D., & Shelton, R.C. (2010). Depressive realism and clinical depression. *Behaviour Research and Therapy* 48, 257–265.
Chater, N. (2018). *The Mind Is Flat: The Remarkable Shallowness of the Improvising Brain*. New Haven, CT: Yale University Press.
Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
---. (2016). *What Kind of Creatures Are We?* New York: Columbia University Press.
---. (2018). Mentality beyond consciousness. In G.D. Caruso (ed.), *Ted Honderich on Consciousness, Determinism, and Humanity* (Cham, Switzerland: Palgrave Macmillan), 33–46.
Chopik, W.J., & Edelstein, R.S. (2014). Death of a salesman: Webpage-based manipulations of mortality salience. *Computers in Human Behavior* 31, 94–99.
Cooper, J. (2005). *Cognitive Dissonance: 50 Years of a Classic Theory*. London: SAGE.

Cooper, J., & Mackie, D. (1983). Cognitive dissonance in an intergroup context.
Corns, J. (2018). Rethinking the negativity bias. *Review of Philosophy and Psychology* 9, 607–625.
Croyle, R.T., & Cooper, J. (1983). Dissonance arousal: Physiological evidence. *Journal of Personality and Social Psychology* 45(4), 782–791.
Cushman, F. (Forthcoming). Rationalization is rational. *Behavioral and Brain Sciences*.
Dacey, M. (2019). Association and mechanisms of priming. *Journal of Cognitive Science* 20(3), 281–321.
D'Cruz, J. (2015). Rationalization as performative pretense. *Philosophical Psychology* 28(7), 980–1000.
Dillon, D.G. (2015). The neuroscience of positive memory deficits in depression. *Frontiers in Psychology* 6(1295), 1–12.
Dogramaci, S. (2014). Intuitions for inferences. *Philosophical Studies* 165(2), 371–399.
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, 280–288.
Elkin, R.A., & Leippe, M.R. (1986). Physiological arousal, dissonance, and attitude change: Evidence for a dissonance-arousal link and a "don't remind me" effect. *Journal of Personality and Social Psychology* 51(1), 55–65.
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), 382–394.
Evans, J.St.B.T., and Stanovich, K.E. (2013.) Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 8(3), 223–241.
Fagerli, R.Aa., & Wandel, M. (1999). Gender differences in opinions and practices with regard to a "healthy diet". *Appetite* 32, 171–190.
Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.
Festinger, L., & Carlsmith, J.M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology* 58(2), 203–210.
Firestone, C., & Scholl, B.J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences* 39, 1–77.
Florian, V., & Mikulincer, M. (1998). Terror management in childhood: Does death conceptualization moderate the effects of mortality salience on acceptance of similar and different others? *Personality and Social Psychology Bulletin* 24(10), 1104–1112.
Fodor, J. (2003). *Hume variations*. Oxford: Clarendon.
Fodor, J.A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition* 28, 3–71.
Freud, S. (1917). Mourning and melancholia. Tr. J. Strachey, in *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (London: Hogarth Press), 243–258.
Friedman, R.S., & Arndt, J. (2005). Reexploring the connection between terror management theory and dissonance theory. *Personality and Social Psychology Bulletin* 31(9), 1217–1225.
Gawronski, B., & Bodenhausen, G.V. (2006). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology* 44, 59–127.
Gilbert, D.T., Pinel, E.C., Wilson, T.D., Blumberg, S.T., & Wheatley, T.P. (1998). Immune neglect:

A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology* 75(3), 617–638.

Gilbert, D.T. (2006). *Stumbling on Happiness*. New York: Vintage Books.

Glass, D.C. (1964). Changes in liking as a means of reducing cognitive discrepancies between self-esteem and aggression. *Journal of Personality* 32(4), 531–549.

Greenberg, J., Martens, A., Jonas, E., Eisenstadt, D., Pyszczynski, T., & Solomon, S. (2003). Psychological defense in anticipation of anxiety: Eliminating the potential for anxiety eliminates the effect of mortality salience on worldview defense. *Psychological Science* 14(5), 516–519.

Hafer, C.L., & Sutton, R. (2016). Belief in a just world. In C. Sabbagh & M. Schmitt (eds.), *Handbook of Social Justice Theory and Research* (New York: Springer), 145–160.

Hamlin, J.K., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science* 13(6), 923–929.

Harman, G. (1986). *Change in View*. Cambridge, MA: MIT Press.

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. *Advances in Experimental Social Psychology* 41, 119–166.

Harmon-Jones, C., Haslam, N., & Bastian, B. (2017). Dissonance reduction in nonhuman animals: Implications for cognitive dissonance theory. *Animal Sentience* 1(12).

Hawks, J., Hunley, K., Lee, S-H., & Wolpoff, M. (2000). Population bottlenecks and Pleistocene human evolution. *Molecular Biology and Evolution* 17(1), 2–22.

Heitland, K., & Bohner, G. (2010). Reducing prejudice via cognitive dissonance: Individual differences in preference for consistency moderate the effects of counter-attitudinal advocacy. *Social Influence* 5(3), 164–181.

Helmholtz, H. (1867). *Handbuch der Physiologischen Optik*.

Holmes, D.S. (1970). Differential change in affective intensity and the forgetting of unpleasant personal experiences. *Journal of Personality and Social Psychology* 15(3), 234–239.

Ito, T.A., & Cacioppo, J.T. (2005). Variations on a human universal: Individual differences in positivity offset and negativity bias. *Cognition & Emotion* 19(1), 1–26.

Ito, T.A., Larsen, J.T., Smith, N.K., & Cacioppo, J.T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology* 75(4), 887–900.

Jenkin, Z. (Forthcoming). The epistemic role of core cognition. *Philosophical Review*.

Johnston, C.S., Krings, F., Maggiori, C., Meier, L.L., & Fiori, M. (2016). Believing in a personal just world helps maintain well-being at work by coloring organizational justice perceptions. *European Journal of Work and Organizational Psychology* 25(6), 945–959.

Jonas, E., Greenberg, J., & Frey, D. (2003). Connecting terror management and dissonance theory: Evidence that mortality salience increases the preference for supporting information after decisions. *Personality and Social Psychology Bulletin* 29(9), 1181–1189.

Jonas, E., Fritsche, I., & Greenberg, J. (2005). Currencies as cultural symbols: An existential psychological perspective on reactions of Germans toward the Euro. *Journal of Economic Psychology* 26, 129–146.

Kahneman, D. (2011). *Thinking Fast and Slow.* New York: Farrar, Straus, and Giroux.
Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist* 39(4), 341–350.
Kant, I. (1781). *Kritik der Reinen Vernunft*. Second edition, 1787.
Killingsworth, M.A., & Gilbert, D.T. (2010). A wandering mind is an unhappy mind. *Science* 330(6006), 932.
Kitayama, S., Chua, H.F., Tompson, S., & Han, S. (2013). Neural mechanisms of dissonance: An fMRI investigation of choice justification. *NeuroImage* 69, 206–212.
Knowles, E.D., & Lowery, B.S. (2012). Meritocracy, self-concerns, and whites' denial of racial inequity. *Self and Identity* 11(2), 202–222.
Knowles, E.D., Lowery, B.S., Chow, R.M., & Unzueta, M.M. (2014). Deny, distance, or dismantle? How white Americans manage a privileged identity. *Perspectives on Psychological Science* 9(6), 594–609.
Lambert, A.J., Eadeh, F.R., Peak, S.A., Scherer, L.D., Schott, J.P., & Slochower, J.M. (2014). Toward a greater understanding of the emotional dynamics of the mortality salience manipulation: Revisiting the "affect-free" claim of terror management research. *Journal of Personality and Social Psychology* 106(5), 655–678.
LeMoult, J., & Gotlib, I.H. (2019). Depression: A cognitive perspective. *Clinical Psychology Review* 69, 51–66.
Lerner, M.J. (1980). *The Belief in a Just World: A Fundamental Delusion.* New York: Springerf
Lerner, M.J. (1965). Evaluation of performance as a function of performer's reward and attractiveness. *Journal of Personality and Social Psychology* 1(4), 355–360.
Lerner, M.J., & Simmons, C.H. (1966). Observer's reaction to the "innocent victim": Compassion or rejection? *Journal of Personality and Social Psychology* 4(2), 203–210.
Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496.
Lieberman, M.D., Ochsner, K.N., Gilbert, D.T., Schachter, D.L. (2001). Do amnesics exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change. *Psychological Science* 12(2), 135–140.
Lipkus, I.M., Dalbert, C., & Siegler, I.C. (1996). The importance of distinguishing the belief ina just world for self versus for others: Implications for psychological well-being. *Personality and Social Psychology Bulletin* 22(7), 666–677.
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, 156–159.
Luka, B.J., & Barsalou, L.W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory & Language* 52, 444–467.
Ma, Q., Meng, L., Wang, L., & Shen, Q. (2014). I endeavor to make it: Effort increases valuation of subsequent monetary reward. *Behavioural Brain Research* 261, 1–7.
Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs* 50(3), 629–658.

---. (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language* 34, 141–157.
Malmgren, A-S. (2018). Varieties of inference? *Philosophical Issues* 28(1), 221–254.
Markie, P. (2017). Rationalism vs. empiricism. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, URL = https://plato.stanford.edu/archives/fall2017/entries/rationalism-empiricism.
McHugh, C., & Way, J. (2018). What is reasoning? *Mind* 127(505), 167–196.
Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* 90(2), 227–234.
Mills, C. (2007). White ignorance. In S. Sullivan & N. Tuana (eds.), *Race and Epistemologies of Ignorance* (Albany: SUNY Press), 13–38.
Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition* 154, 40–48.
Montoya, R.M., Horton, R.S., Vevea, J.L., Citkowicz, M., & Lauber, E.A. (2017). A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological Bulletin* 143(5), 459–498.
Moore, M.T., & Fresco, D.M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychological Review* 32, 496–509.
Nelson, J.C., Adams, G., & Salter, P.S. (2013). The Marley hypothesis: Denial of racism reflects ignorance of history. *Psychological Science* 24(2), 213–218.
Neta, R. (2013). What is an inference? *Philosophical Issues* 23, 388–407.
Nichols, S., Strohminger, N., Rai, A., & Garfield, J. (2018). Death and the self. *Cognitive Science* 42, 314–332.
Norton, M.I., Mochon, D., & Ariely, D. (2012). The IKEA effect: When labor leads to love. *Journal of Consumer Psychology* 22(3), 453–460.
Olson, M.A., & Fazio, R.H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science* 12(5), 413–417.
Orlandi, N. (2014). *The Innocent Eye*. New York: OUP.
Piazza, J., Ruby, M.B., Loughnan, S., Luong, M., Kulik, J., Watkins, H.M., & Seigerman, M. (2015). Rationalizing meat consumption: The 4Ns. *Appetite* 91, 114–128.
Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. London: Viking.
Poerio, G.L., Totterdell, P., & Miles, E. (2013). Mind-wandering and negative mood: Does one thing really lead to another? *Consciousness and Cognition* 22, 1412–1421.
Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.
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.
Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies* 175, 2353–2372.
Reed, A. (2014). We are all right-wingers now: How Fox News, ineffective liberals, corporate Dems and GOP money captured everything. *Salon*. Interviewed by T. Frank. https://www.salon.com/2014/03/09/we_are_all_right_wingers_now_how_fox_news_ineffect

ive_liberals_corporate_dems_and_gop_money_captured_everything/, accessed 26 November 2019.

Ritter, C., Benson, D.E., & Snyder, C. (1990). Belief in a just world and depression. *Sociological Perspectives* 33(2), 235–252.

Rothgerber, H. (2013). Real men don't eat (vegetable) quiche: Masculinity and the justification of meat consumption. *Psychology of Men & Masculinity* 14(4), 363–375.

---. (2014). Efforts to reduce vegetarian-induced dissonance among meat eaters. *Appetite* 79, 32–41.

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), 629–643.

Rozin, P., & Royzman, E.B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review* 5(4), 296–320.

Ruby, M.B., & Heine, S.J. (2011). Meat, morals, and masculinity. *Appetite* 56, 447–450.

Sætrevik, B., and Sjåstad, H. (unpublished). A pre-registered attempt to replicate the mortality salience effect in traditional and novel measures. Accessed on PsyArXiv, 20 June 2019, DOI: 10.31234/osf.io/dkg53.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Science* 69(5), 379–399.

Schaffner, B.F., & Luks, S. (2018). Misinformation or expressive responding? What an inauguration crowd can tell us about the source of political misinformation in surveys. *Public Opinion Quarterly* 82(1), 135–147.

Schimel, J., Hayes, J., Williams, T., & Jahrig, J. (2007). Is death really the worm at the core? Converging evidence that worldview threat increases death-thought accessibility. *Journal of Personality and Social Psychology* 92(5), 789–803.

Schopenhauer, A. (1851). *Parerga and Paralipomena, Volume 2*. Tr. E.F.J. Payne, 1974 (Oxford: OUP).

Schwitzgebel, E., & Ellis, J. (2017). Rationalization in moral and philosophical thought. In J.-F. Bonnefon & B. Trémolière (eds.), *Moral Inferences* (Routledge), 170 –190.

Shea, N. & Frith, C. (2016). Dual-process theories and consciousness: The case for 'Type Zero' cognition. *Neuroscience of Consciousness*. doi: 10.1093/nc/niw005.

Sherratt, K.A.L., & MacLeod, A.K. (2013). Underlying motivation in the approach and avoidance goals of depressed and non-depressed individuals. *Cognition & Emotion* 27(8), 1432–1440.

Siegel, S. (2017). *The Rationality of Perception*. Oxford: OUP.

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

Stanovich, K.E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning* 24(4), 423–444.

Steele, C.M. Southwick, L.L., & Critchlow, B. (1981). Dissonance and alcohol: Drinking your troubles away. *Journal of Personality and Social Psychology* 41(5), 831–846.

Stone, J., & Cooper, J. (2003). The effect of self-attribute relevance on how self-esteem moderates attitude change in dissonance processes. *Journal of Experimental Social Psychology* 39, 508–515.

Taylor, S.E. (1991). Asymmetrical effects of positive and negative events: The mobilization–minimization hypothesis. *Psychological Bulletin* 110(1), 67–85.
Valaris, M. (2014). Reasoning and regress. *Mind* 123(489), 101–127.
Velleman, D. (2000). On the aim of belief. In his *The Possibility of Practical Reason* (New York: OUP), 244–281.
Walker, W.R., Vogl, R.J., & Thompson, C.P. (1997). Autobiographical memory: Unpleasantness fades faster than pleasantness over time. *Applied Cognitive Psychology* 11, 399–413.
Walker, W.R., Skowronski, J.J., Gibbons, J.A., Vogl, R.J., & Thompson, C.P. (2003). On the emotions that accompany autobiographical memories: Dysphoria disrupts the fading affect bias. *Cognition and Emotion* 17(5), 703–723.
Wang, C.E., Brennen, T., & Holte, A. (2006). Decreased approach motivation in depression. *Scandinavian Journal of Psychology* 47(6), 505–511.
Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavioral Research* 45, 1191–1207.
Watkins, P.C., Vache, K., Verney, S.P., Muller, S., & Mathews, A. (1996). Unconscious mood-congruent memory bias in depression. *Journal of Abnormal Psychology* 105(1), 34–41.
Wilkins, C., & Kaiser, C.R. (2014). Racial progress as threat to the status hierarchy: Implications for perceptions of anti-white bias. *Psychological Science* 25(2), 439–446.
Williamson, T. (2000). *Knowledge and Its Limits*. Oxford: OUP.
Yon, D., Bunce, C., & Press, C. (2020). Illusions of control without delusions of grandeur. *Cognition* 205, 104429.
Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology* 9(2), 1–27.
Zanna, M.P., & Aziza, C. (1976). On the interaction of repression-sensitization and attention in resolving cognitive dissonance. *Journal of Personality* 44(4), 577–593.
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.
Zhao, W.J., Walasek, L., & Bhatia, S. (2020). Psychological mechanisms of loss aversion: A drift-diffusion decomposition. *Cognitive Psychology* 123, 101331.