Précis of *Talking to Our Selves: Reflection, Ignorance, and Agency*

John M. Doris
Philosophy-Neuroscience-Psychology Program, Philosophy Department, Washington University in St. Louis, St. Louis, MO 63130
jdoris@wustl.edu
http://www.moralpsychology.net/jdoris/

Abstract: Does it make sense for people to hold one another responsible for what they do, as happens in countless social interactions every day? One of the most unsettling lessons from recent psychological research is that people are routinely mistaken about the origins of their behavior. Yet philosophical orthodoxy holds that the exercise of morally responsible agency typically requires accurate self-awareness. If the orthodoxy is right, and the psychology is to be believed, people characteristically fail to meet the standards of morally responsible agency, and we are faced with the possibility of skepticism about agency. Unlike many philosophers, I accept the unsettling lesson from psychology. I insist, however, that we are not driven to skepticism. Instead, we should reject the requirement of accurate self-awareness for morally responsible agency. In *Talking to Our Selves* I develop a *dialogic* theory, where the exercise of morally responsible agency emerges through a collaborative conversational process by which human beings, although afflicted with a remarkable degree of self-ignorance, are able to realize their values in their lives.

Keywords: agency; character; confabulation; deliberation; ethics; morality; psychology; reasoning; reflection; responsibility; the self; self-knowledge; value

If you haven't already despaired of politics, consider Ballot Order Effects: candidates topping the slate may enjoy a several point advantage in vote share (Krosnick et al. 2004, pp. 61–68; cf. Lutz 2010; Marcinkiewicz 2014; Meredith & Salant 2013; Webber et al. 2014). The data don't reveal the thinking of individual voters, but I doubt it's anything like this:

> *I'll vote for her because she's first on the ballot.*

Political discourse ranges shamelessly over the daft and deplorable. But you don't often hear a voting rationale like *that*.

Ballot Order Effects illustrate the phenomena animating *Talking to Our Selves* (Doris 2015b),[1] which I call *incongruent parallel processing* ("incongruence" for short), where two (or more) cognitive processes (with "cognitive" understood capaciously) deliver divergent outputs regarding the same object (pp. 51–52). I interpret incongruence under the rubric of "dual process" theories that are currently ubiquitous across the sciences of mind (pp. 49–51). The approach has its critics, and the details are debated (Evans & Frankish 2009), but my purposes require only the broadest – and least doctrinaire – of brushes. On one familiar characterization, automatic processing is supposed to be effortless (sometimes "mandatory"), fast, and associated with emotional and other functioning exhibiting comparatively little cognitive elaboration, while analytic processing is supposed to be effortful (sometimes "discretionary"), slow, and associated with "higher" cognition (e.g., Stanovich 2004, pp. 37–47; Wilson 2002, pp. 52–53). Analytic processing, on such accounts, supports ratiocination of the sort celebrated in philosophy – reflection about what to think and do, and why to think and do it (cf. Frankish & Evans 2009, p. 15), while the "quick and dirty" of automatic processing seems scarcely worthy of the honorific *reasoning* (p. 50).

Many of the most philosophically trenchant instances of incongruence occur when the automatic bests the analytic and prompts people to do things they wouldn't endorse (or do things from motives they wouldn't endorse). Here, the *causes* of a person's behavior would not be regarded by her as *justifying reasons* for that behavior (pp. 43–44, pp. 64–65). That is, she would be unwilling to cite in defense of her behavior the factors figuring in the most

John M. Doris is a Professor in the Philosophy–Neuroscience–Psychology Program and Philosophy Department at Washington University in St. Louis. He is the author of *Lack of Character: Personality and Moral Behavior* (Cambridge, 2002) and *Talking to Our Selves: Reflection, Ignorance, and Agency* (Oxford, 2015). With his colleagues in the Moral Psychology Research Group, he wrote and edited *The Moral Psychology Handbook* (Oxford, 2010). Doris has been awarded fellowships from Michigan's Institute for the Humanities, Princeton's University Center for Human Values, the National Humanities Center, the American Council of Learned Societies, the Center for Advanced Study in the Behavioral Sciences, and the National Endowment for the Humanities. He is a winner of the Society for Philosophy and Psychology's Stanton Prize for excellence in interdisciplinary research.

© Cambridge University Press 2018
0140-525X/18

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

perspicuous psychological explanation of her behavior, were she aware of these factors (as she very often might not be). In such cases, the person's reasoning is somehow *bypassed* (p. 52; Nahmias 2011, pp. 560–563): if your vote gets decided by ballot order, it seems as though your preferences and judgment – assuming you're not of the improbable opinion that your vote *should* be so determined – have been left out of the decision.

Bypassing raises doubts about the extent to which human beings exercise *morally responsible agency* – roughly, the way human beings order their own behavior in a fashion that merits the distinctively ethical responses they direct at one another (pp. ix, 7, 23–33). Admiring the kindness of your friend is different from appreciating the beauty of a rainbow; your friend's act of kindness is a *doing*, or something they accomplished, while the emergence of a spine-tingling rainbow is a *happening* for which nobody is due credit (unless you're inclined to credit divine agency). I want to make good theoretical sense of this difference; human beings exercise morally responsible agency, while less intelligent natural systems like rainbows don't. The occurrence of bypassings intimate that human beings are not, contrary to appearances, in fact so distinguished – or at least not so distinguished as one might have wished.

In considering the variety of disconcerting phenomena, like Ballot Order Effects, that intimate the existence of bypassing incongruence, *Talking to Our Selves* articulates, and then attempts to ameliorate, *skepticism* about morally responsible agency. Skepticisms about agency sourced in empirical psychology have been stated – and perhaps overstated – before (e.g., Wegner 2002), and these skepticisms themselves invite a healthy skepticism (e.g., Bloom 2014). Nevertheless, I agree with the skeptics, in as much as I think that when the science is properly appreciated, there is cause for deep concern about the prospects for morally responsible agency. This appreciation makes the work of the book's first part. But I also agree with the anti-skeptics, because I think there's a good answer to the skeptical provocation. This answer requires walking some unfamiliar paths in thinking about agency and moral responsibility, and that walking makes the work of the book's second part. To answer the skeptic, I try to establish that (1) human beings exercise morally responsible agency with tolerable regularity, and (2) have epistemic resources to identify such exercises in everyday moral discourse and practice.

1. Preliminaries

To avoid confusion, an initial terminological orientation is required (p. 23, n. 4). As I've indicated, the theory of agency I'm after is a theory of *morally responsible* agency, a notion rather more demanding than thinner senses of agency, where "agent" may reference any entity capable of purposive movement. This thinner sense distinguishes agents like puppies from objects like Pop-Tarts, and is common in psychology, as with the developmental literature. For example, human infants behave in ways indicating that they may categorize novel objects – in one study, an entity poetically titled an "agentive blob" – as agents, if the objects display performances indicative of perception, communication, and goal-directedness (Johnson 2003).

The properties sufficient for being an agentive blob are insufficient for being a morally responsible agent, even if they do mark important differences between puppies and Pop-Tarts. Lots of critters – honeybees, for example – exhibit perception, communication, and goal-directedness, without engendering serious temptation toward subjecting them to morally important responses; fine to be angry *that* you were stung by a bee (one damn thing after another!), but inapposite to be angry *at* the bee (which is, after all, only a bee). I've no interest in disparaging the sort of agency displayed by agentive blobs and honeybees – an important fact, if detecting this kind of agency is prominent in the human cognitive repertoire. It is not, however, sufficient for morally responsible agency; that notion, as I try to show, requires rather more. While I sometimes drop the modifier "morally responsible," it is, unless otherwise noted, a morally responsible agency at issue whenever "agency" appears here (and in the book).

At this juncture, I should also provide methodological orientation, and say something about the status of the empirical material on which my remorselessly interdisciplinary approach relies. As the twenty-first century sputters toward its third decade, numerous replication failures attending iconic studies, particularly in experimental social psychology, have – as they should – occasioned much soul-searching by producers and consumers of the social sciences, and have – as they should – encouraged greater methodological scrupulosity in social scientific practice. The "RepliGate" controversy is not yet concluded (and involves controversy about whether there should even be a controversy[2]), which means that *Talking to Our Selves* includes some of my own soul-searching about interpreting science in conditions of uncertainty (pp. 44–49; cf. Machery & Doris 2017).

RepliGate, and the inevitability of scientific controversy more generally, dictate taking one's science with a healthy dose of caution. But it should be stressed that empirical findings intimating the existence of incongruence are not "one off" curiosities, but are part of established trends, and the dual process perspective in which I situate the phenomena have been proposed for most everything psychologists study: attention (Schneider & Shiffrin 1977), learning (Reber 1993), memory (Roediger 1990), perception (Norman 2002), reasoning (Evans & Over 1996), decision making (Kahneman & Frederick 2002), person interpretation (Gilbert et al. 1988), delay of gratification (Metcalfe & Mischel 1999), psychopathology (Beevers 2005), and moral judgment (Cushman et al. 2010).

The theoretical division presupposed by dual process theory doubtless is, like all theoretical divisions, rather less crystalline in world than on page (p. 50). (It should especially be noted that the analytic/automatic distinction crosscuts the intelligent/unintelligent distinction: it's not smart to concoct intricate excuses for procrastination, and it's not dumb to reflexively bolt from danger.) Nevertheless, that different cognitive processes can proceed independently of, and sometimes oppositionally to, one another is a hypothesis supported by an impressive range of converging evidence (pp. 49, 52, 56). Indeed, even critics of dual process theory (e.g., Machery 2009, pp. 121–150; Keren & Schul 2009, pp. 141–42) acknowledge the existence of the cognitive conflict I'm calling incongruent parallel processing.

While individual studies ought to be critically examined on their merits (pp. 53–64), the skeptical worry does not depend on the fate of particular "criterial" experiments. (If the Ballot Order Effect turned out to be chimerical, for example, I'd have plenty of other illustrative effects to

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

choose from.) Whatever controversy afflicts particular studies, labs, or research programs, the existence of incongruence (however exactly it should be theorized) isn't in question. It is this *un*controversial observation on which my argument relies.

2. Skepticism

I presume that the problem animating *Talking to Our Selves* is "intuitive"; I presume many folks will think that incongruence sits uneasily with their (pre-theoretical) self-understanding (pp. x, 80, 158). That is, I presume that for those people who believe that they're directing their own behavior and shaping their own life (as I presume many people believe), learning that they were prey to something like a Ballot Order Effect would give them serious pause.

Say you are an undecided voter, and as you enter the booth, you decide, finally, to help make your country great again, and vote for Candidate Creepy. Say that subsequently, I provide convincing evidence that you in fact voted for Creepy because his leading place on the ballot tipped your decision – if not for ballot order, you'd have voted for his slightly less objectionable rival. (For the moment, don't mind that the actual evidence is aggregate, and can't decide individual cases; more on that in a bit). How should you feel about your decision?

This pre-theoretical unease, I further contend, also afflicts many philosophical theories of agency and responsibility, such as the influential approaches associating agency with "reasons responsiveness" (Fischer & Ravizza 1998; Nelkin 2011; Vargas 2013). Suppose you read research associating sedentary lifestyles with poor health outcomes, and commence dragging yourself to the gym; your behavior is responsive to the balance of reasons commending it, and you've thereby, with your exercising, pulled off an exercise of agency. But for your (imaginary) vote, you took yourself to be responding to Creepy-reasons, like reclaiming lost greatness, while your vote was actually decided by a rationally arbitrary "non-reason," Creepy's place on the ballot. Here, your conduct was not appropriately responsive to reasons, and therefore not properly agential, according to reasons responsiveness theories. My point is not that these theories – or any other philosophical theory – *cannot* accommodate the disconcerting facts, but that they *need* to do so (p. 171). An anti-skeptical theory of human agency – one maintaining that exercises of morally responsible agency may be confidently identified in a good measure of human behavior – must account, in a systematic and detailed way, for the phenomena of incongruence, but philosophical efforts in this direction have hitherto been limited.

As I've said, the provocative cases of incongruence involve psychological causes of behavior that are not plausibly taken as reasons for that behavior, as when a "dumb" automatic process bypasses a "smart" analytic one. Theorizing reasons is the stuff of uncounted dissertations in philosophy and the decision sciences, but one doesn't need fancy theory to appreciate the difficulty. A simple, broadly subjective, account here serves: when someone treats a consideration as a reason they should be willing to treat that consideration as a *justification* for their judgment or behavior (pp. 43–44). The notion of justification is also a dissertation bogey, but once again, an elaborate theory isn't needed (nor need one be attributed to the reasoner). Many people commend and defend what they think and do, both to themselves and others, and when they engage in this activity, they are engaged in a reason-giving, justificatory practice (p. 44). The relevant divergence of reasons and causes obtains when people would decline to reference the psychological origins of their behavior – those origins that would figure in a reasonably complete and accurate psychological explanation of that behavior – as appropriate considerations in this practice.

I don't primarily intend familiar cases of what philosophers call *akrasia*, or weakness of will, where people succumb to desire, appetite, or temptation against their better judgment (pp. 52, 161–162): I know the sugar, grease, and gluten-filled donut is a nutritional apocalypse I'll surely regret later, but I nonetheless devour it like a starving lion. However much trouble they make for health and happiness, many such cases don't make special trouble for agency. For when I inhale the donut, I'm doing so on perfectly intelligible grounds: Winslow's salt-caramel donuts are absurdly tasty, I know I'll quite enjoy eating one, and I very much desire to do so. Nor is it the case that the desire is somehow puzzling, or alien to me; gustatory pleasure is something that may be appropriately treated as a reason, and I sometimes treat it that way. Much as I hate to admit it, I suffer an altogether unpuzzling affection for Winslow's salt-caramel donuts, and if this affection manifests in my body composition, it manifests as a result of my exercising agency, and is nobody's fault but mine. (Were I *addicted* to Winslow's salt-caramel donuts, there would be questions about my agency, but my interest here is the ordinary, not the psychopathological.)

In the most unsettling cases of incongruence, judgment is not overwhelmed, as in *akrasia*, but bypassed. And in many such cases there is little reason to think reason would be overmatched were the bypassing factor subject to scrutiny: probably not unduly difficult to disregard the influence of the Ballot Order Effect, when it's brought to one's attention. That's part of what makes the phenomena so unnerving. Hardly surprising that people can be moved by strong desires; more so, that they can be moved by factors that are scarcely intelligible as the objects of desire (p. 162).

A preliminary schema for skepticism about morally responsible agency is now visible (pp. 64–65). Cases of incongruence, where the pertinent psychological causes of her cognition or behavior would not be recognized[3] by the actor as reasons favoring that cognition or behavior (were she aware of these causes at the time of performance), are cases where *defeaters* obtain. Where defeaters obtain, the exercise of agency does not obtain. If the presence of defeaters cannot be confidently ruled out, it is not justified to attribute the actor an exercise of agency. If there is general difficulty in ruling out defeaters, skepticism about agency ensues. (Again, *akratic* desirers will not always be defeaters, because such desires are frequently recognized as reasons [pp. 70–71].) In brief, the agency skeptic maintains that for any putative instance of agency, one cannot rule out the presence of a defeater (or defeaters) in the causal origins of that behavior, and one is therefore never justified in positing an instance of morally responsible agency.

Here, the standard for out-ruling is substantially ethical (p. 66). Responsibility attributions are associated with

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

distribution of benefits (like reward and praise) and burdens (like punishment and blame), so the improper attribution of responsibility can result in the target of the attribution receiving unjust benefits or burdens. Thus, the requirement that the possibility of defeaters be "confidently" eliminated gets understood in terms that are more moral than purely epistemic. For one to be justified in an attribution of morally responsible agency, it should be the case that if one's initial attribution of agency is overturned by credible new evidence of defeaters, one is not guilty of wrongdoing for having apportioned benefit or burden according to the initial attribution. Viewed in this light, the skeptical thought is that the requisite moral confidence is never justified; one cannot be sufficiently confident that one is not guilty of wrongdoing when acting on attributions of morally responsible agency. So understood, the skeptic is not making an unreasonable demand for epistemic certainty (the familiar skeptical foible of "setting the bar too high"), but an eminently reasonable request that moral judgments be morally defensible.

It must be emphatically emphasized that the skeptical argument (like other skeptical arguments) does *not* require an empirical generalization about the frequency with which defeaters occur (p. 68). And a good thing: that argument is sure to be a stinker, whatever side one is on. For so far as I can see, nobody has any very exact idea what the frequencies are. It's obvious why: comparatively little behavior is closely observed, still less behavior is observed in controlled conditions, and for the behavior that is observed, there's often little assurance about the full range of explanatorily salient psychological antecedents. If the contest between skepticism and anti-skepticism had to be decided by speculation about frequencies, it wouldn't get decided.

Nevertheless, if the skeptical argument is to be more than a symptom of philosophical paranoia, the possibility of defeaters has to be something more than *mere* possibility. And so it is. The skeptical hypothesis – concerning the pervasive availability of behavioral explanations referencing defeaters – is not a loopy (and perhaps massively unlikely) proposition like philosophical fantasies of Evil Demons or Envatted Brains. Rather, it is a "live" hypothesis (Frances 2005, pp. 560–61): it has been vetted by the relevant experts, and judged by a substantial number of them, on the basis of good evidence, to be about as likely as competing hypotheses (p. 66; cf. Davies 2009, p. 169). Given the repeated appearance of incongruence in the context of scientific research, it is responsibly conjectured that defeaters occur with some regularity in everyday life; regularly enough, anyway, that the skeptical hypothesis is live, and has some non-trivial chance of being true, for any particular behavior. The matter can't be put more precisely than that. Nor need it be: the critical question concerns not how often defeaters should be thought to obtain, but how their presence can be ruled out for each putative exercise of morally responsible agency (p. 68).

I suspect anti-skeptical optimism is commonly funded by the imprecise suspicion that the envisaged defeaters are "pretty rare," together with the sanguine assurance that pretty rare occurrences need not undermine the attribution of m orally responsible agency. Regarding the sanguine assurance, I'm not sure: depending on the gloss of rare or pretty rare, such events can certainly merit moral concern. ("Better to acquit 1,000 guilty persons than to execute a single innocent one.") I suppose there is some gloss that warrants the assurance – 1 in 100?, 1,000?, 10,000? – but such a gloss is unavailable, faced with inevitable uncertainty about frequencies. And the imprecise suspicion fares no better than the sanguine assurance. Given how little is observed of – and the less that is known about – human psychology and behavior, it is obscure what could motivate justifiable confidence that defeaters are pretty rare. At least, that motivation had better not be (as I fear it often is) the unvarnished faith that the counterintuitive must be uncommon – a certitude inimical to the spirit of scientific discovery.

This framing allows us to better understand the relation of aggregate evidence, such as Ballot Order Effects, to particular instances, such as your vote for Creepy. What the aggregate indicates is that *some* individuals must be subject to the influence in question; otherwise, there would not be an effect (pp. 63–64). We cannot, for studies of this kind, be sure *which* individuals are affected. But this is not an *objection* to the argument, it *is* the argument. Once it be allowed that defeaters might be in play – and that's what must be allowed – we require considerations sufficient to warrant moral confidence they are not. That's the burden of ruling out. As with other skeptical arguments, the present skeptical argument is not *undermined* by uncertainty; it *trades* in uncertainty.

It should be noted that many of the discomfiting experimental findings may involve small effect sizes. One way to consider this question is in terms of the correlation coefficient (r), which quantifies the strength of association (covariation) between two variables: a coefficient of 1.0 indicates a perfect association, a coefficient of 0.00 indicates no association, and a coefficient of −1.0 indicates a perfect negative association. Effect size for correlation coefficients may be assessed with reference to a venerable proposal by Cohen (1988, pp. 77–81) that for psychology (at least in "softer" branches like personality and social) a correlation of around 0.10 should be regarded as "small," around 0.30 as "moderate," and around 0.50 as "large."

Alas, for much psychological research, correlation coefficients rating Cohen's large, or even moderate, occur relatively infrequently (Hemphill 2003). When the famous "group effect" (Darley & Latané 1968) is calculated as a correlation between the number of onlookers and bystander intervention, the value is a "moderate" −.38 (Funder & Ozer 1983, p. 110), and many celebrated findings in psychology, such as those relating implicit and explicit bias, may involve considerably smaller effects (Greenwald et al. 2015). Are small effects "too small"? (Cohen [1988, p. 79] himself didn't think so.) For context, consider a rendering of effects in biomedical research as correlation coefficients (Meyer et al. 2001, p. 130): regular aspirin consumption and reduced risk of heart attack, 0.02; chemotherapy and surviving breast cancer, 0.03; ever smoking and lung cancer within 25 years, 0.08. These relationships are not relationships, like the correlation between being human and being mortal, that are strong enough to be detected by "the naked eye" without the aid of statistical magnification (Jennings et al. 1982, pp. 216–22). And obviously, they are clearly small by Cohen's lights.

However, such small effects may yet have practical import: take the aspirin, endure the chemo, and don't light up. The small relationship between aspirin use and

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

reduced coronary risk might be thought of as pretty important, compared to something that (presumably) has no relationship, like owning a gray coat, or something that has a negative relationship, like obesity. Of course, if your habits or genes are bad enough, aspirin ain't gonna save you. Yet for many patients the costs associated with taking an aspirin every day or two are pretty minimal, so if you're at risk for a heart attack, why not? Furthermore, taking aspirin together with other interventions, like losing weight and exercising properly, might have a considerable cumulative effect.

If we can say that some lives are saved by small-effect interventions like aspirin (perhaps in concert with other interventions), why shouldn't we say that some exercises of agency are undermined by small-effect defeaters, like Ballot Order Effects (perhaps in concert with other defeaters)? To be sure, identifying statistically small aggregate effects does not allow confident conclusions about particular outcomes for particular individuals. But the small effects must be making a difference in *some* individual cases, or there would not be aggregate effects. Given the multitude of influences likely operative in any instance, one cannot confidently say where the difference was made, but this sort of uncertainty is actually part of the problem: *what's* making *which* difference for *whom*?

Once we allow that there are some of these rationally and ethically arbitrary influences on cognition and behavior, we are bound to admit there may be others. If stuff like *that* can make a difference, there could be *many* such influences in any particular instance. And while the impact of each such individual influence may be statistically small, the cumulative effect may be quite potent. For all one knows, any decision may be infested with any number of arbitrary influences. The claim is not that any one of the influences in question is momentous in the way illness, bereavement, and unemployment can be. Rather, the thought is that statistically small effects can sometimes be practically consequential – and an aggregation of such influences even more so.

I'm compelled to admit that I previously, while espousing skepticism about traditional conceptions of character, complained about small effect sizes in personality psychology, and it's arguable that some effects in personality are larger, perhaps considerably larger, than the kinds of effects I'm now celebrating. But for character skepticism, my concern was not that the personality effects are small, but that they are likely to be *smaller than should be expected* on familiar theories of character and personality (Doris 2002, pp. 68, 71–75). Here, my position is the converse: the effects in question, small though they may sometimes be, are *larger than should be expected* on familiar theories of agency and responsibility – indeed, it often seems absurd that they could have *any* effect at all. This forces rethinking approaches to agency.

3. Reflectivism

Like much philosophy, *Talking to Our Selves* is structured agonistically. My primary agonist is *reflectivism* (pp. x, 17–23), a doctrine according to which *the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do*. Typically, this doctrine is associated with a corollary: *the exercise of human agency requires accurate reflection*. In an exercise of agency, as construed by reflectivism, a person correctly divines the beliefs, desires, and other psychological states relevant to her decision, makes her decision in light of these states, and then acts accordingly. In short, reflectivism holds that the exercise of agency is characteristically reflective activity.

Reflectivism is another notion I guess is pretty darn "intuitive," in the sense of enshrining a very familiar experience: you think about what you're hungry for, in perusing the menu, and I think about how much space I need, in perusing the real estate section. When we have these experiences of reflective agency, I confidently speculate, we don't usually think the outcome of our reflections was determined – to tweak my stock example – by the order of listings on the menu or in the adverts. (*Let's rent the more expensive flat, it's listed first!*)

I also insist that reflectivism is philosophically familiar: preoccupation with reflection is, arguably, the Western philosophical tradition's most distinctive feature, in both historical and contemporary guises, and is certainly a central theme in philosophical moral psychology (Doris 2015b, p. 17; see also Arpaly 2002, p. 20; Kornblith 2010, p. 2; 2012, p. 1). The most salient examples are the many Kantians in moral psychology and ethics, who often associate practical rationality with reflection (Korsgaard 1996, pp. 92–93; 2009, p. xi; cf. Moran 2001, p. 127), resulting in a literature thick with references to the "reflective agent" and "reflective agency" (Velleman 1989, p. 5; 2000, pp. 12, 26–29, 124, 191–96; Wallace 2003, p. 437; 2006, pp. 150–51).

The accuracy corollary is equally recognizable (pp. 19–20). For reflectivists, agency requires that the actor detect practically salient facts about herself (Velleman 2000, p. 12); as Tiberius (2002, p. 13) puts it, a "person who does not know her deepest motivations because she is very psychologically complicated, or because she has formed impenetrable layers of self-deception, is missing something she would need to deliberate well." If deliberation – which might be thought of as practical reflection – is predicated on erroneous self-understandings, agency is supposedly imperiled.

On this understanding, the self-ignorant don't deliberate effectively about what to think and do; in contrast, the practically effective "reflective agent" will enjoy some measure of accuracy in her deliberations. When the reflectivist's reflective agent cites reasons for her action, the reasons cited are supposed to accord with the causes of her action: if she believes she did something because she judged it the right thing to do, that judgment must figure in an accurate and appropriate causal explanation of why she did it.

Unfortunately, the empirical record intimates that reflective activity is probably neither so characteristic of, nor practically important to, human beings as reflectivists suppose it is (p. 22). Many human behaviors are thoughtless, and unconstrained by the deliverances of reflection; on those instances when people do reflect, there is little warrant for confidence that these reflections are informed by accurate self-awareness. If so, there's something seriously wrong with both reflectivism and its corollary (p. *x*), namely, that if reflectivism is the only philosophically viable account of agency, and the best empirical guess is that reflective agency is not an especially prominent form of human self-direction, there's going to be a shortage of morally responsible agency (p. 152).

The reflectivist might not be dismayed at this prospect; the exercise of agency is an achievement, and it might be that exercises of reflective agency are a scarce achievement (pp. 35–37). Or perhaps reflectivism trades in ideal agency rather than actual agency, and eschews empirical claims (pp. 22, 151). These are defensible positions, but perhaps of dubious appeal. My sensibilities in this respect are conservative (p. 158): I suppose that people (at least in the cultural context where this discussion resides) regularly attribute moral responsibility to one another (whether tacitly or explicitly), and I further suppose that enactments of this familiar practice are not pervasively mistaken. That is, I presume treating the intentional behaviors of normal healthy adults as exercises of morally responsible agency is a morally defensible "default" (pp. 33, 39). So I think the envisaged agency shortage would be the worse for reflectivism, not for the practice (p. 33). But deciding exactly how frequent exercises of agency are, on reflectivism or any other theory, is to enter a kind of discussion I've already said is doomed. Instead, the difficulty is how to rule out the existence of defeaters, and I don't think extant reflectivisms are well situated to do so.

Philosophical targets are famously elusive, and philosophers routinely disavow the views attributed them by critics (pp. 17, 108). (*That's not my position, and anyway, your objection to it doesn't work.*) It's therefore likely that some reflectivists will deny being committed to views compromised by the arguments I derive from the empirical literature. I disagree, and cite textual evidence for my reading (pp. 17–21) – which ascribes a familiar and intuitive view to my agonists – but this isn't the kind of disagreement that is likely to get amicably resolved.

Rather than playing too long at pin-the-tail-on-the philosopher, it is better to proffer a friendly challenge. Reflectivists typically do not consider in detail how to accommodate the empirical difficulties associated with incongruence (pp. 164, 171), and their theorizing about agency would benefit if they did so. The critical, "targeting" part of the book may be understood not as excoriation, but as *invitation* – an invitation to help advance the cause of empirically credible theorizing in moral psychology. The positive part of the book can be seen as an attempt to identify one path to such advance, *via* departing reflectivism. Because I endeavor to avoid unseemly triumphalism (p. 171), I don't deny that there may also be reflectivist means of egress. At this stage of debate, however, it is unclear to me how those means will be developed, or how successful they are likely to be.

4. Experience

Maybe the best "ruling-out" response to skepticism is plain experience (p. 78), in particular the experience of agency: the feeling of doing something (like deciding what to have for dinner) is manifestly distinct from the feeling of having something happen to you (like dinner setting off your digestion). I don't deny such experiences are commonplace, and I further grant that many "agent experiences" (my unlovely name for the feeling of doing) bear a reflectivist character, where it seems to one that one self-consciously and accurately inventories one's inclinations and circumstances, decides what to do on the basis of this inventory, and acts according to one's decision (p. 80).

What I do deny is that agent experiences have the stuff to block skepticism, because such experiences are surprisingly untrustworthy (pp. 96–97). (This is not to say that people are wrong about the experience of agency in the sense of recognizing when their movements are endogenously rather than exogenously generated [Gallagher 2000]; again, at issue is the more demanding notion of *morally responsible* agency.) My denial is steeped in the clinical literature on *confabulations*, the extravagant fabrications produced by patients suffering neurological trauma or psychiatric illness (pp. 81–89). Given many striking demonstrations of failed self-awareness in healthy people, it's commonly suggested that the unafflicted confabulate, much as the afflicted do. The parallels are imperfect, but it is the case that healthy people suffer substantial self-ignorance, and that first-personal reports of self-awareness are of questionable reliability.

Likely implicated in such failures is the well-documented class of phenomena united under the heading "motivated cognition" (pp. 94–96). Although its prevalence is a matter of controversy, and the boundaries between motivational and cognitive processes are muddy, there seems little doubt that motivation powerfully influences cognition, so that reasoning and belief may be determined by extra-epistemological factors (Dawson et al. 2002, pp. 1379–81; Ditto & Lopez 1992; Dunning et al. 1999, p. 79; Gilovich 1991, pp. 75–87, 84; Kruglanski 1996; Kunda 1990, p. 493).

For my purposes, perhaps the most salient variety of motivated cognition is self-enhancement (pp. 92–94): People are prone to inflated assessments of their attributes and performances (Alicke et al. 2001, p. 9; Dufner et al. 2012, p. 538; Dunning 1999, pp. 5–6; Dunning 2006). For example, undergraduates may think themselves more popular than they really are (Zuckerman & Jost 2001), while their professors may think themselves better teachers than they really are (Blackburn & Clark 1975, p. 249; cf. Cross 1977).

I interpret the experience of agency as a kind of self-enhancement: people believe they control things they do not, a tendency called the "illusion of control" (pp. 134–36). For example, gamblers commonly seem to think they influence chance events (Davis et al. 2000, pp. 1236–37; Toneatto et al. 1997, p. 262; cf. Langer 1975): in lotteries around the world, quick pick options, where a computer picks random numbers, account for only 10–20% of tickets sold (Simon 1998, p. 247). As one player insisted, you shouldn't "trust the computer to pick your numbers. I trust myself more" (as quoted in Farrell et al. 2005, p. 597).

It's not just that people exaggerate their control of their circumstances; they overestimate their control of their selves. Here's a patient with hemiplegic anosognosia "explaining" her inability to move her arm (p. 87): "I have never been very ambidextrous"; "I've got severe arthritis in my shoulder"; "Doctor, these medical students have been prodding me all day and I'm sick of it. I don't want to use my left arm" (as quoted in Ramachandran 1996, p. 125). The last is especially suggestive, as it replaces patiency with agency: For a person suddenly experiencing the helplessness of the sick role, the performance can be understood as a kind of self-enhancement, albeit a tragically ineffectual one.

I suppose that at least some such confabulations are sincere (p. 84); the anosognosic is not necessarily

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

prevaricating. But whether or not this patient was experiencing agency, her words are – significantly – a self-presentation of agency. As we shall see in a moment, this kind of agential self-presentation is a central feature of human social life.

5. Values

Under the most generic description, my understanding of agency – ponderously titled an *anti-reflectivist*, *valuational*, and *dialogic* (or *collaborativist*) theory – represents a philosophically commonplace (though not universally endorsed) "compatibilism" (pp. 9–12): I think that moral responsibility is compatible with causal determinism, and people may sometimes be held morally responsible in circumstances where their behavior is determined by factors "external" to themselves. For compatibilists, the important question is not *that* a behavior was caused but *how* it was caused: for exercises of morally responsible agency, behavior must be caused *in the right way*.

Like many other compatibilists, I make a start on "in the right way" with P. F. Strawson's "reactive attitudes": the motley of emotional and other interpersonal responses – indignation, anger, gratitude, admiration, and the like – with which human beings regulate their social lives (p. 23; Strawson 1962; Vargas 2004; 2008; 2013; Watson 1993). For me, thinking about what (if any) reactive attitudes are apt helps establish whether someone is morally responsible (or not): once again, I shouldn't be angry at the bee who stings me, but it's perfectly appropriate for me to be angry at you if you poke me with a fork in a fit of dinner-hour pique. The reason for the differing reactions, my story goes, is that you are morally responsible for your behavior, and the bee not. The difference in responsibility, as my story goes on, is explained by the observation that you were exercising agency, and the bee not. The aptness of moral responsibility attributions and the associated reactive attitudes are, for behaviors of moral concern, characteristic symptoms of agency.

Next, I say a behavior is an exercise of morally responsible agency when the actor is *self-directed* while performing it, and further assert that behavior is self-directed when it expresses the actor's values (after Watson 1975; 1996; cf. Bratman 2007, p. 48; Smith 2005; Sripada 2015a; 2015b). While this particular valuational theory is only one of numerous contending theories, and hardly the object of philosophical consensus, it does enable some pretty plausible observations: when the nicotine addict guiltily succumbs to craving and lights up, his behavior is not self-directed, but when he manages to resist a craving because he values his health, his behavior is self-directed.

For me then, morally responsible agency gets understood in terms of self-direction, and self-direction gets understood in terms of expressing values: behaviors are exercises of agency when they are expressions of the actor's values. In turn, values get understood in terms of desires (Bratman 2007, pp. 47–67; Harman 2000, p. 135) – desires possessed of a reasonably reliable and substantial motivational force. Not just any such desires: the desires properly associated with value are those desires the actor accepts in a determinative role for her practical planning (Bratman 2007, pp. 64–66). For a desire to be associated with a value, it must also have a justificatory role: it must be something that the planner is amenable to employing in justification or defense of her plan.

The planner, to my way of thinking, need not be aware of her willingness to assign this justificatory role (pp. 27–28). People may have desires, values, and plans that they are quite unaware of, and their behavior may express their values without their knowing that it does so. (Consult locutions like, "I guess that was my plan all along," and "I suppose this must be what I really wanted.") This is a crucial feature of my account, because it makes room for agency in the absence of accurate reflection. If I speak out of cruelty while thinking I'm speaking out of honesty, you may well hold me responsible for hurting your feelings. And if pressed to defend your attribution, you might say something like, "Whatever he thought he was doing, he really *wanted* to be cruel." Pretty good reason to attribute an exercise of agency, where someone acts on what matters to her. The less likely thing would be to say that someone did what mattered to her, because it mattered to her, but didn't do so as a morally responsible agent (pp. 160–61).

This provides occasion to make a bit clearer what the argument is *not* about (p. 69). It's commonly supposed that much behavior lacking "conscious control" makes trouble for agency (e.g., Levy 2011, pp. 188–94; Wegner 2002, pp. 156–58, 170–71; 2005, p. 28). But this supposition immediately encounters inconvenience, for it apparently excludes far too much from the offices of agency. Acts of kindness or callousness that are done habitually, with little in the way of conscious supervision, are customarily treated, and perfectly appropriately so, as exercises of morally responsible agency. The point generalizes, and widely. While there's controversy concerning the role of conscious control in skilled behavior (Christensen et al. 2016; Montero 2010; Noë 2012, Ch. 6), surely many skilled behaviors are performed unreflectively (how else could they be done so quickly and effortlessly?), and just as surely, people are often credited – appropriately so – for these performances.

Adopting a valuational theory of agency like the one I adopt is not necessarily to reject reflectivism; a valuational theory might be fashioned as a reflectivist theory, as is, perhaps, Bratman's (2007 p. 28, n. 5). Rather, the dispute concerns what psychological processes *facilitate the exercise of agency*. The reflectivist taps self-conscious, tolerably accurate reflection in this role, while I do not require accurate reflection for the exercise of agency. And this difference, I claim, leaves me better situated than the reflectivist to ameliorate agency skepticism.

6. Ignorance

Ignorance remains a paradigmatic excuse, which may block the attribution of responsibility (without Horton's big ears, the other critters couldn't have known that they were endangering the dust speck world of Whoville). But *self*-ignorance, such as self-enhancement and illusions of control, doesn't necessarily undermine morally responsible agency, and may often *enable* it (pp. 129, 136, 144, 158).

One important pathway by which self-ignorance may facilitate agency is *motivational* (pp. 136–37, 144). Turns out, writing this précis was nearly impossible for me, which seems a bit odd: why should it be hard to summarize

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

a book you've already finished? A likely explanation notes that the book is muddled, and the author of modest ability. But this path leads to despair; if I take the elucidation seriously, I'll abandon my labors for the gym, a nap, or worse. On the other hand, if I attribute my troubles in summarizing to the profundity of my topic and the intricacy of my thought, I may be hopeful enough to fight, and write, another day. The point extends: plodding Professor Drudge might better be able to grind out long hours at his desk under the auspices of "talented" than "mediocre," and given the contribution of perspiration to inspiration, this understanding can help Drudge produce work of sufficient quality to earn him the title "talented" – his mediocrity notwithstanding (p. 144). If Drudge values his professional status, his positivity may facilitate his agency.

Another example: unrealistic positivity has been implicated in improved health outcomes (Taylor et al. 2003), and it appears that this self-enhancement extends to perceptions of control. In a study of cancer patients, perceptions of control were negatively associated with maladjustment; patients with higher perceptions of control were less likely to experience anxiety and depression (Thompson et al. 1993, pp. 297–298). Moreover, there was a larger negative relationship between perceptions of control and maladjustment for those rated lower in physical functionality (Thompson et al. 1993, pp. 299–300); perceived control may have been doing more good for those who had less control of their circumstances. Given the abundant evidence implicating psychological well-being in physical well-being, and psychological distress in physical distress (Brenner 1979; Diener & Chan 2011; Steptoe et al. 2005; Veenhoven 1988), it becomes very tempting to say that elevated perceptions of control, in so far as they're associated with better adjustment, are good for one's health.

Sick or well, valuing my health is presumably part of the reason I strive to promote it. Perhaps my beliefs about the efficacy of these efforts are motivated cognitions; I may think my labors more effective than they are, in part because I want them to succeed. But these motivat*ed* cognitions may simultaneously be motivat*ing* cognitions; my believing my efforts will work inspires me to undertake work that in fact work, even if they don't work so well as I believe (pp. 135–37). Here, self-ignorance, *via* motivational pathways, promotes the exercise of agency.

A similar story can be told for romantic relationships (136–37). According to one group of marital researchers (Fowers et al 2001, pp. 96–99, 102, 105), the presence of positive illusions is "nearly universal" among "satisfied spouses"; the mean estimate of divorce given by members of married couples was 10%, and the modal estimate was zero%, while scientific estimates of its likelihood are often in the range of 40–60% (Fowers et al. 2001, p. 105; for more on "marital optimism," see Baker & Emery 1993; Boyer-Pennington et al. 2001). Hand wringing over the "50% divorce rate" is, as any reader of *Divorce Magazine* (2004) can tell you, a prominent component of public discourse in the United States; apparently, romantic illusions may persist in the face a well-known body of undermining fact (difficulty in estimating divorce rates duly noted).

At the same time, relationship outcomes may be sensitive to effort: numerous studies indicate that couples counseling is effective (Bray & Jouriles 1995; Hahlweg & Richtera 2010; Sayers et al. 1998; Snyder et al. 2006). During relationship trouble, I'm willing to speculate, sturdy perceptions of control have motivational utility: if you don't think the quality of your relationship is responsive to effort, how do you get yourself to undertake the effort? But if you think couples counseling can help you walk it back from the brink, could be you, yours, and your therapist will put the needed work in.

By supporting value-conducive motives, illusions of control may facilitate behavior that helps realize values. Self-ignorance often functions to effect self-direction, and its absence can be an impediment to agency: a complete and accurate understanding of your career, health, or relationship prospects might prevent you from making them all you want them to be. If so, we've identified a pathway whereby self-ignorance supports, rather than impairs, agency. This pathway is often indirect (p. 127): falsely believing I can directly effect a valued outcome may support motivation eventuating in behaviors or circumstances that do, in fact, effect the outcome in question. The fact, we can say, is the child of the fiction (pp. 136–37).

7. Collaboration

To fill out this story, I understand self-ignorance in the context of sociality. In doing so, I identify a secondary agonist, *individualism*, which maintains that optimal human reasoning is exemplified by individual thinkers (pp. 103, 107–109). Against this, I pit *collaborativism* about rationality, where optimal human reasoning is held to be "socially embedded," and then extend this collaborativism to agency, where many important exercises of *individual* agency are substantially *social* phenomena (pp. 103, 115, 122).

Start with simple examples (pp. 123–24): seatbelt laws have been found to increase seatbelt use (Gantz & Henkle 2002; Shults et al. 2004), while public health campaigns have been found to decrease rates of smoking (Fiore et al. 2000). Assuming that people value their lives and health, these are cases where social processes enable people to better express their values in their conduct. But while there is little doubt that postponing death and disability is (for most people) a valued outcome, one might wonder if realizing these outcomes should count as an exercise of agency. The mechanism matters, and succumbing to media manipulation or yielding to government coercion may not seem appropriately agential.

Perhaps psychotherapy makes a more convincing illustration (pp. 124–25). People *seek out* psychotherapy in hopes of making their lives go better, so this *active* process has a more agential appearance than the succumbing to media campaigns or state regulations. Of course, people often enter therapy in response to something like duress: a stalled career, a strife-torn marriage, or stacks of unpaid bills. But these incentives are not obviously inimical to agency. On the contrary, people trying to change their lives for the better is an excellent place to look for agency, and psychotherapy can help effect this change.

Outcome studies, using both clinician assessment and client self-report, indicate that talk therapy works – it can ameliorate various adverse psychological conditions, such as depression and anxiety (Lambert & Ogles 2004; Luborsky et al. 1985, p. 609; Seligman 1993). We've now an appealing example of collaborativism about agency. First,

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

the "talking cure" is very much a social treatment, where client and therapist work things through together more effectively than the client could do on their own. There's the collaborativism. Second, decreasing psychological discomfort and increasing personal efficacy are very likely values many clients in therapy hold, so the clinical process is reasonably thought to facilitate the expression of these values. There's the agency.

I contend that the success of this endeavor does not require accurate self-awareness on the part of the client. (Or, perhaps, accurate awareness of the client by the therapist; interestingly, the rubric under which therapy is conducted is not a critical determinant of clinical efficacy [Brown et al. 1999; Dawes 1994, pp. 38–74; Luborsky & Singer 1975; Wampold et al. 1997; Woolfolk 1998].) At the same time, a recurring theme in the clinical literature is that a "positive alliance" between therapist and client is associated with successful outcomes (Horvath & Symonds 1991; Krupnick et al. 1996; Martin et al. 2000; Orlinsky et al. 2004). For many consumers of psychotherapy, I'm guessing that comes as a relief: if you always end up talking to your therapist about fluff like television or sports instead of the deepest workings of your soul, you may yet be doing yourself good, so long as you're bonding with your therapist.[4] Supposing therapy can facilitate agency, as I've just suggested, we've here a case where agency is achieved without accurate self-awareness. In this instance, collaborativism and anti-reflectivism are complementary.

8. Rationalization

To better understand the synergistic contribution of self-ignorance and collaboration to morally responsible agency, consider Johansson, Hall, and colleagues' incredible studies of "choice blindness" (Hall et al. 2010; Johansson et al. 2006), where people fluently provide agential explanations *for choices they didn't make* (pp. 138–40). In Sweden, Hall et al. (2012; cf. Hall et al. 2013) demonstrated choice blindness for moral and political attitudes. People strolling through a park were given a twelve-item survey with statements concerning either general moral principles or current moral issues, and asked to report their attitudes on a 9-point scale anchored at "completely agree" and "completely disagree." After completing the survey, participants read aloud three of the statements they had responded to, and explained their positions. In manipulated trials, two of these statements were reversed: if someone originally agreed with "Even if an action might harm the innocent, *it can still be morally permissible to perform it*," it now appeared that they had agreed to "If an action might harm the innocent, then *it is not morally permissible to perform it*" (emphasis added).

The reversal was noticed in only 47% of trails, and 69% of participants accepted at least one reversed statement. Although the politically active were more likely to correct reversals, people claiming to generally hold strong moral opinions weren't more likely to make corrections. Unsurprisingly, level of agreement was associated with correction: The more participants agreed or disagreed with a statement, the more likely they were to correct the reversal. But nearly a third (31.4%) of all manipulated trials with answers at the endpoints of the scale (1 or 9) were not corrected. And remarkably, when it came to explaining manipulated choices, 53% of the participants *argued unequivocally for the reversal of their original position*. People are able to quite assuredly justify and explain their choices – even when "their" choices are not choices they made!

While folks sometimes hold their nose and knowingly "go along to get along," the social influence in choice blindness studies most probably proceeds subliminally. Otherwise it would be difficult to explain the fluidity with which participants explained their (non-) choices (p. 138; Johansson et al. 2006), as well as the apparently genuine surprise participants evince in post-experimental debriefing (Hall et al. 2010). But while the pseudo-explanations weren't conscious social niceties, they have something importantly in common with social niceties: such explanations are required by convention. It can be socially awkward to stand mute when questioned about the reasons for one's political or moral convictions; commonly, keeping silent or pleading ignorance won't do, and most any answer, or at least a wide range of answers, is better than not answering (Hirstein 2005, pp. 4–5).

Not so much for garden-variety factual ignorance; fine to say that I don't remember the name of the 23rd U.S. president, or the first woman to summit Everest. But where reasons are required, not so much. I suspect that instances of "rational dumbfounding" (pp. 140–41) where people are unable to explain their behavior – particularly their intentional behavior – are pretty unusual. Even if rational dumbfounding is more common than I suspect, *confessing* rational dumbfounding still seems remarkable. People fluently produce socially serviceable explanations for what they do, even when their self-ignorance extends to the psychological origins of their behavior.

I call these performances *rationalizations* rather than confabulations, to distinguish them from clinical confabulation, and I deploy "rationalization" in a non-pejorative sense, absent any connotations of bad faith (pp. 141–43). Here, a rationalization is a (typically verbal) performance that presents judgment and behavior as rational. (Or, slightly less circularly, rationalizations make judgment and behavior make sense; cf. Gibbard 1990, pp. 37–38, 156–59). If I'm right, a central form of rationalization presents a behavior as an exercise of agency: *I chose to do so. I meant to do that. I had my reasons.* Frequently, rationalizations may reflect illusions of control: people present themselves as achieving exercises of agency even when they have not.

My account favors approaches that construe agency as structured by narrative (Doris 2015b, pp. 143–46; see also Dennett 1991; 1992; Fischer 2006, pp. 106–23; 2009, pp. 145–77; Schechtman 1996 2011; Velleman 2006; *pace* Strawson 2004). Who people are, and what they do, is shaped by the self-depictions, which I call biographies, they express to themselves and others. That one understands one has *made* a promise, and further understands oneself to be a person who *honors* her promises, may help ensure that the promise is *kept*. One's biography can secure behavior expressing one's values, even in the face of unfavorable circumstance or instable inclination. Thinking – and talking – of oneself as a dutiful promiser may be causally implicated in one keeping promises on those occasions that one does so, whether one is a dutiful promiser or not. Then to do its work biography needn't be accurate, so long as it is motivationally engaging for the teller.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Biographies may be private; I might live by a story I tell only to myself. But biographies are also presented socially, and serve as vehicles for the exchange of rationalizations. It's because I've had the life I've had – or seem to have had – that I'm justified in doing as I do. The trauma of experiencing a near-fatal automobile accident, for example, may justify my disinclination to see a movie prominently featuring car chases, and prompt you to propose seeing a less bombastic film. Human beings develop rationalizations collaboratively, and the central requirement for these rationalizations is not accuracy, but accord with one's interlocutors. People shape their lives, not as isolated reflectors, but as participants in an ongoing negotiation – a negotiation that simultaneously constrains and expresses who they are. In a slogan, agents are negotiations. If you like, call this notion of agency *dialogic* (p. 148). Here morally responsible agency requires not *freedom* from influence, but *mutual* influence, as individuals express their values in a collaborative process.

9. Skepticism (again)

I have an account of how morally responsible agency may be exercised, but ameliorating the skeptical problem also requires an account of how exercises of morally responsible agency are detected (pp. 159–64). The skeptic says we are never justified in attributing moral responsibility; in retort, the anti-skeptic must articulate conditions when it is justifiable to do so. (Reminder: the relevant notion of justification is substantially ethical.)

According to my valuational approach, archetypal exercises of morally responsible agency are expressions of the actor's values, so the problem is determining whether the actor's values are expressed in their conduct. To do so, on my understanding of the skeptical challenge, we must be justified in ruling out the presence of defeaters; when we have done so, we may have the requisite moral confidence that the relevant conduct expresses the actor's values, and may therefore be justified in attributing an exercise of morally responsible agency (p. 159).

In preparation for this work, we ought to realize that valuing has a temporal dimension: values are expressed over time, and often can only be identified over time (pp. 162–63). It will frequently be difficult to determine whether someone holds a value, in the absence of temporally extended trends in cognition, rationalization, and behavior. But with extended observation, a pattern of symptoms may emerge: as cognitions, rationalizations, and behaviors appropriate to a value tend to recur in a person's life, we, and they, may begin to have confidence that a person holds that value, and that particular behaviors, patterns of behavior, or life projects are expressing it (even where their behavior is less than consistent with respect to that value, as it very probably will be, if "situationism" about moral personality and behavior [Doris 2002] is anything close to correct).

Conversely, if one focuses on isolated events, diagnosis may falter. (The same is true of medical diagnosis; that's why your doctor takes your history, or holds you overnight for observation.) It will frequently be obscure whether someone doing something is an expression of her values; the evaluative signal may be quite weak, against the background of situational noise.

In tracking the evaluative signal, sociality will be central; first-personal inquiry will frequently be augmented by second personal, collaborative, inquiry: a friend observes that I'm often downcast after a day at my "dream job," or an old lover remarks that you seem much happier with your new partner (p. 163). Collaboration also occurs in institutional contexts: therapeutic and educational endeavors can help people figure out what matters to them. Sometimes, others have better access – or at least instructively different access – to a person's values than does the person herself. The extent of agency-impairing self-ignorance is considerable, but people are collectively possessed of epistemic assets fit to ameliorate it, assets deployed in the continuing social negotiation by which people order and make sense of their lives.

Attribution of agency and responsibility may be warranted when a pattern of cognition, rationalization, and behavior emerges, and that pattern is best explained as involving the expression of some value. Determining whether a particular action expresses a value, in the sense of being governed by a value relevant goal, as opposed to fortuitously conforming to that value (pp. 25–26), will very frequently make difficult work. But the emergence of trends across iterated cognitions and behaviors can underwrite confidence that the trend is to be accounted for by reference to a person's values, rather than a massively coincidental run of defeaters. Typically, the required evidence base must be both wide, covering multiple observations of behavior, and deep, licensing inference to the psychological states implicated in the behaviors. In such cases, the presence of defeaters need not be treated as a live possibility, which means that in such cases the skeptical challenge is defanged.[5]

These welcome conditions may obtain less often than one might wish; given the vagaries of mind and world, defensibly attributing exercises of morally responsible agency, on any plausible theory, takes hard work. But if one focuses on isolated cases of reflective deliberation, as the reflectivist is wont to do, the work is much harder (p. 164). There, the possibility of defeaters – given all that is known about the potential for rationally arbitrary influences on judgment and decision – cannot usually be ruled out with confidence sufficient to warrant attribution of responsibility. Reflectivist paradigms of morally responsible agency, then, are epistemically challenged.

Reflectivists may insist they can accommodate my advice to depart emphasis on atomistic behaviors in favor of extended processes. Fair enough. But there's not only the problem of how agency is identified, there's also the problem of how agency is facilitated (p. 164). A compelling response to the skeptic will provide standards for attribution of agency, *and* an account of how agency may be realized in human lives – lives afflicted with surprisingly high levels of self-ignorance. Articulating standards will be cold comfort, without an account of how people may live up to those standards. A dialogic understanding of agency offers one such account – an explanation of how agency emerges in the face of limited self-awareness, through a process of collaborative negotiation. If accurate self-conscious reflection is required for morally responsible agency, the prospects for agency look rather worse.

Supposing I've now decent answers to the problems of how exercises of morally responsible agency are facilitated and identified, another question remains (pp. 164–66):

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

does my approach to responsibility capture the "normative" character of responsibility related discourse and practice? It's widely accepted that normativity has something to do with the guidance of thought and behavior: as opposed to descriptive questions about how the world is, normative questions are prescriptive questions concerning what ought be done about it. Normative discourse, then, is *oughty* discourse. If so, a theory of responsibility should explain why attributions of responsibility (and their denial) carry imperatives about how the subjects of such attributions should be regarded and treated.

Accounts of responsibility such as mine, centered on reactive attitudes, capture something of normativity rather easily. The various reactive attitudes may make a motley assortment, but whatever else they are, they often involve emotions. Emotions are standardly thought to involve "action tendencies" (Nichols 2007, pp. 412–414), and even where they don't immediately move people to action, they prepare people for action: emotions structure the range of behavioral options (Prinz 2004, pp. 191–96).

To suffer an emotion is to be told what to do, or not do. But emotional imperatives don't carry all of the normativity that might be desired. For people can, and do, ask whether their emotions are appropriate, justified, or fitting (D'Arms & Jacobson 2000; 2006). In so doing, they're asking about what is helpfully called normative authority (Railton 2003, p. 344): why should a command issued by emotion command my assent?

When responsibility is understood as I understand it, by way of reactive attitudes, the challenge is to identify compelling theoretical grounds for when and what reactive attitudes are appropriate. On my theory, responsibility is associated with the exercise of agency, and the exercise of agency with expressions of values, so the question becomes whether these expressions are appropriate targets for the reactive attitudes.

I think this question is readily answered: when someone's deeds manifest their values, it makes good sense to direct anger or admiration their way. I'm angry with the Wall Street Oligarch who orders a million-dollar renovation for his office ($1,400 wastebasket included) as his company fails and the economy falters, because I think he values status too much and humanity too little. I admire the man who donates 10% of his $1,000 monthly disability check to charity, because I think he's moved by the opposite complex of values. And were I pressed to justify my reactions, I could make a convincing case on the grounds of what matters, and fails to matter, to each man. The point might be put in the language of desert (Doris 2015a; Vargas 2013, pp. 234–66): reference to each man's values explains why they deserve the attitudes I subject them to.

In associating responsibility attribution with emotionally infused reactive attitudes, I find something of the oughtiness associated with normativity. And by locating agency in the expression of values, I've located a perspicuous rationale for these reactive attitudes: there's pretty good reason for you to be angry with me for what I did, if what I did is a function of my mean-spirited matterings. This, it seems to me, is an account possessed of sufficient normative authority for the discourse and practice of responsibility attribution.

I don't insist my way is the only way. Because I am a *pluralist* about agency and responsibility, I allow that morally responsible agency may be exercised in other ways (pp. 171–77). But one of the ways people sometimes exercise agency, the way envisaged by the reflectivist, has been seriously overemphasized by philosophers, with the result that many philosophical theories are poorly situated to accommodate incongruence.

10. Selves

If one is overly ambitious, as I confess to being, one can extend valuational and dialogic perspectives on morally responsible agency to two notorious "problems of the self" (pp. 5–9, 179–97): continuity – *what is required for a person to survive changes?* – and identity, *what distinguishes one person from another?* (Philosophers use "identity" in both of these contexts; I've departed this dual usage by using "continuity" in the context of survival.) Actually, joining these two problems to the problem of agency is, however ambitious it may be, altogether necessary. For if persons are not entities persisting over time while possessed of relatively determinate identities, it is obscure on what responsibility is to affix.

To my thinking, as the problems are related, as is the solution: in completing my theory, I extend the collaborativist, valuational, dialogic approach to continuity and identity. I begin this extension by proposing that continuity is *socially contingent* (p. 182): personal continuity is predicated on psychological continuity, and psychological continuity is sustained by societal continuity, so if a perturbation in circumstances is substantial enough, as in cases of cultural devastation like those associated with the North American genocide (pp. 178–80), personal continuity may be compromised, physical survival notwithstanding.

My proposal seemingly conflicts with a standard dictum in the philosophical literature on personal identity: *extrinsic factors don't count* (p. 182). According to this dictum, whether a person at one time bears an "identity relationship" to a person at another depends only on facts about the two "person stages" and the relations between them (Noonan 1989, p. 152). Your survival has to do with *you*, and not your neighbors, friends, or family: a neighbor dying, or even a lot of neighbors dying, doesn't mean you've failed to survive.

Yet a moment's thought reveals that *lots* of stuff – within your skin and without – has to do with you and your survival. Given that persons are entwined in causal webs with strands trailing far beyond their body, it's hard to say what's intrinsic to a person and what's extrinsic. It's nevertheless obvious that cultural conditions are causally related – powerfully and pervasively so – with the psychology of those associated with them. One can therefore conclude that culture may affect personal continuity, and is not excluded from consideration by any plausible "no extrinsic factors" principle. Even if you're more confident than I in the existence of a tenable extrinsic/intrinsic divide, you shouldn't want to deny that identity-intrinsic factors may be causally impacted by social and cultural perturbations.

Here's one avenue for such impact: *personal continuity varies with evaluative continuity* (p. 183). When a person's values change, they become less like the person they were before; at the extreme, if a person's values are completely changed, I contend, they are no longer the same person. Culture is an important determinant of value, and cultures ebb and flow; so too does value.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Typically, in cultural change, circumstances are mixed: some practices go on, and others falter; some values persist, and others fade away. But it remains the case that when sufficient cultural upheaval occurs, there may be disruptions of the practices required to support central values that are substantial enough to press questions about psychological continuity – and thereby personal survival. Answers to these questions will seldom be all or nothing; although change is a constant, significant evaluative continuity is likely the rule rather than the exception. Yet this point seems secure: factors well beyond the skin matter for continuity.

Explicit discussions of identity are less visible in mainstream "analytic" philosophy than discussions of continuity, but we can make a start with Taylor (1994, p. 25), who describes identity as "something like a person's understanding of who they are, of their fundamental defining characteristics as a human being." Terms like "fundamental" and "defining" are, unfortunately, not easily subdued. However, a hopeful opening gambit proceeds in terms of *individuation* (p. 187): one's defining characteristics are the constellation of attributes by virtue of which one is different from other people – even while one shares many attributes with other people (Appiah 2010, Ch. 1; Taylor 1994, p. 28).

I understand individuation, as I understand agency and continuity, by reference to value (p. 188). Identity may be a source of value: a ritual observance might be of value to me by virtue of my identity as a member of a certain religion, but valueless to those who do not share my faith. Identity may also be the object of value: someone might value her profession, ethnicity, or political affiliation. Most important here is that value may be a source of identity (the relation between identity and value is bi-directional). A person has the identity they do partly by virtue of having the values they do: Who I am has much to do with what matters to me.

If I'm right, a dialogic valuational theory can be crafted into a comprehensive theory of the self, because it accounts for notions central to philosophical thinking on selves (pp. 6–9): agency, continuity, and identity. The theory can also explain how agency, continuity, and identity emerge in actual human lives, by way of an ongoing process of collaborative rationalization. This process develops and sustains the self, despite the limits of reflection and the infirmities of self-awareness. It also provides material by which to answer skepticism about morally responsible agency, because the collaborative process facilitates identifying a person's values, and the role of these values in their behavior. That's a lot to get done! Hopefully, *Talking to Our Selves* makes a decent start on doing it.[6]

NOTES

[1] From here on, *Talking to Our Selves* is generally referenced parenthetically, by page number only.

[2] For some recent contributions, ranging over the alarmed to the sanguine, see Anderson et al. (2016), Gilbert et al. (2016), Inbar (2016), Open Science Collaboration (2015), and Van Bavel et al. (2016).

[3] One might wish to insert a "good faith" rider: On (too) many occasions, people deny treating things as reasons that they do in fact treat as such.

[4] A regret: in hindsight, I should have thought harder about the contribution of emotion to the exercise of agency, a topic that deserves much more discussion than it typically receives.

[5] Defanged assuming "fallibilism," where the required justification is not *conclusive*, but *defeasible* (65).

[6] Many thanks to Miranda Alperstein, Justin D'Arms, Dan Haybron, Edouard Machery, Shaun Nichols, Laura Niemi, Casey O'Callaghan, Paul Bloom, and Manuel Vargas for their generous help on earlier drafts. I'm especially appreciate Julia Staffel's encouragement and comments on two previous versions. Writing was completed during a term as a Laurence S. Rockefeller Fellow at Princeton's University Center for Human Values. I'm most grateful to the Center, and to Washington University in St. Louis for sabbatical leave.

Open Peer Commentary

The Nietzschean precedent for anti-reflective, dialogical agency

doi:10.1017/S0140525X17000620, e37

Mark Alfano
Ethics & Philosophy of Technology, Delft University of Technology, Delft, Netherlands; Institute for Religion and Critical Inquiry, Australian Catholic University, Sydney, Australia.
mark.alfano@gmail.com www.alfanophilosophy.com

Abstract: Nietzsche anticipates both the anti-reflective and the dialogical aspects of Doris's theory of agency. Nietzsche's doctrine of will to power presupposes that agency does not require reflection but emerges from interacting drives, affects, and emotions. Furthermore, Nietzsche identifies two channels through which dialogical processes of person-formation flow: sometimes a person announces what she is and meets with social acceptance of that claim; sometimes someone else announces what the person is, and she accepts the attribution.

John Doris and Friedrich Nietzsche have a lot in common. In addition to being provocative and humorous writers in their native idioms, they share a conception of human agency. It can be tiresome to point out the priority claims of an earlier philosopher, so I should say at the outset that I do so not to smugly insist that my guy got there first but to showcase a closely allied perspective that may shed additional light and offer glimpses around blind corners. In particular, I argue that Nietzsche anticipates both the anti-reflective and the dialogical aspects of Doris's theory of agency.

Doris's primary target is *reflectivism*, according to which the exercise of human agency consists in judgment and behavior ordered by (accurate enough) reflection about what to think and do. As Paul Katsafanas (2013; 2016) and I (Alfano 2010; 2013b; 2016b) have argued, Nietzsche's doctrine of will to power presupposes that human agency does not require reflection. Instead, agency emerges from the interaction of drives, affects, and emotions that are sometimes in harmony and sometimes in conflict (*Daybreak*, Nietzsche 1881/1997, sect. 119; *Gay Science*, Nietzsche 1882/2001, sects. 333, 354, 357; *Beyond Good and Evil*, Nietzsche 1886/2001, sects. 6, 12, 200, 224; *Genealogy of Morals* I:13, II:16, Nietzsche 1887/2006).[1] While affects and emotions receive much attention in contemporary dual-process psychology, drives – which Nietzsche construes as standing motivational dispositions to engage in particular action-types – have largely been ignored. The Nietzschean perspective thus expands the class of mental processes that can lead to incongruent parallel processing, though empirical research is of course needed to determine the role played by drives in our mental economies.

Like Doris, Nietzsche is concerned that incongruent parallel processing may undermine agency. Doris focuses on cases in

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

which someone would not endorse the causes of her own behavior because those causes are what I have dubbed "non-reasons" (Alfano 2013a, pp. 43–45). While Nietzsche does not rule out such arational influences on behavior, his concern is directed to cases in which someone would outright reject an accurate description of the causes of her behavior (*Genealogy of Morals* I:10–11, II:11, III:15). Nietzsche's solution to the problem of incongruent parallel processing also resembles Doris's. According to Doris, someone's behavior constitutes an exercise of agency just in case it expresses a subset of the agent's values: namely, those values that are sufficiently longstanding, strong, and accepted by the agent as justificatory. The key move in this account is to allow for agency despite self-ignorance. The agent needn't know that she is willing to assign a justificatory role to the values she expresses in action, nor need she realize that her exercise of agency expresses one or more of her values. Similarly, for Nietzsche (*Daybreak*, sect. 109; *Beyond Good and Evil*, sect. 3), someone's behavior constitutes an exercise of agency just in case it expresses a subset of the agent's drives, affects, and emotions: namely, those that, were the agent to learn about them, would not lead her to disapprove of her own action (Katsafanas 2013, p. 138; 2016, Ch. 7). Like Doris, then, Nietzsche allows for agency despite substantial self-ignorance. The agent needn't know that she would endorse her own action after learning more about its etiology, nor need she know which drives, affects, and emotions led to her action in the first place.

Lowering the bar of accurate reflection in this way enables Doris and Nietzsche to countenance and even make use of a common human foible: the tendency to *post hoc* confabulation. When we reflect on why we did what we did or what sorts of people we are, we often enough tell ourselves flattering stories that enhance our own rationality, moral rectitude, or agency – turning every "it was" into a "thus I willed it" (*Thus Spoke Zarathustra* II, Redemption, Nietzsche 1883/2006; see also *Gay Science*, sect. 277). Without disputing the inaccuracy of such confabulation, Doris argues that it can lend someone the courage to move confidently into the future, to undertake ambitious projects, and to make demanding commitments. Tactically deployed fictions about ourselves can become facts. Indeed, in Alfano (2013a, Ch. 4; see also Alfano 2016a; 2016c, Ch. 4) I argue that they can lead to factitious traits of character, where someone becomes, for example, honest because she first falsely attributes honesty to herself.

To understand how people manage this trick, we must turn to the dialogical aspect of Doris's and Nietzsche's conception of agency. Doris's secondary target is individualism, according to which optimal human decision -making is exemplified by individual thinkers. He argues instead for collaborativism, according to which human decision-making and agency are socially embedded and perhaps even extended through dialogical processes. The factitiously honest person is (typically) not honest in splendid isolation. Instead, she engages in social interactions that "hold" her in a network of narratives that represent what she and those close to her consider her most important actions, passions, traits, roles, relationships, and values (cf. Lindemann 2014).

There are several ways in which such narratives help build and stabilize human persons. They can change someone's self-concept by representing them as embodying values or drives that they previously did not self-attribute. They can reassure someone of the accuracy of their self-concept by providing social proof. They can lend someone a communal identity by representing them as one of "us." And they can help "shape and crystallize" someone's traits and values by "making more determinate tendencies and impulses [. . .] that are in some degree inchoate" (Wong 2006, p. 136). If this is right, then talking about our mental lives resembles not so much describing the weather as negotiating a cooperative agreement.

As I argue in Alfano (2015; 2016b), Nietzsche identifies two main channels through which such dialogical processes flow (*Human, All Too Human*, Nietzsche 1878/1996, sect. 51; *Daybreak*, sects. 105, 201, 248; *Gay Science*, sects. 21, 40, 58; *Beyond Good and Evil*, sects. 42, 44, 261; *Genealogy of Morals* I:2, I:6). On the one hand, sometimes a person announces what she is (i.e., what her values, motives, concerns, or drives are), and that announcement meets with social acceptance; on the other hand, sometimes someone else announces what the person is, and she accepts the attribution. Such bid-and-accept patterns can be iterated. X could describe herself has embodying value V, to which Y responds by pointing to evidence (e.g., in her past behavior) that she actually embodies value V*, to which X responds by pointing to evidence that she actually embodies V†, and so on. Moreover, the negotiation needn't be so explicit. X could instead tell a story that represents herself as embodying V, to which Y responds by asking a question that presupposes that she embodies V*, to which X responds by telling another story that represents herself as embodying V†. The kind of person or self that emerges from such feedback loops is reflected or echoed rather than reflective and transparent.

If this is right, then much of human agency is constitutively social. Moreover, it suggests that a novel class of dispositions – namely, the dispositions associated with being a good echo – must be recognized and theorized by philosophers and social scientists.[2] A virtuous echoer may not be entirely accurate and comprehensive. Instead, I contend, a virtuous echoer filters and perhaps even distorts to some extent; in this way, the echoer modulates existing dispositions without inventing them whole cloth. By contrast, a vicious echoer may negotiate their partner into accepting an unflattering self-description or even an incoherent self-description, undermining their agency.

I conclude with a speculation about the future of dialogical agency in the wake of self-tracking technologies and predictive analytics (cf. Selke 2016). If it's true that someone's agency and self are constructed by the stories she tells and is told about herself, and that these stories need be neither wholly accurate nor unfiltered, then such technologies and analytics may curtail one's capacity to – in Nietzsche's phrase – become what one is. Self-tracking is liable to make it more difficult to enjoy self-enhancing illusions. If I tell a story about how I once did this or saw that, it may become possible for me or someone else to verify or falsify my narrative. In addition, predictive analytics is liable to make it more difficult to have inflated confidence in one's capabilities. If positive illusions about my own potential enable me to undertake ambitious plans and commitments, then well-evidenced predictions that I am likely to fail may keep from trying in the first place. There may be a steep prudential price to be paid for the epistemic benefits of these innovations.

NOTES

[1] I refer to Nietzsche's texts using the canonical section numbering rather than page numbers.

[2] Arguably, the literature on transactive memories (e.g., Dixon & Gould 1996) already does this to some extent. Thanks to Alessandra Tanesini for pointing this out.

Innate valuation, existential framing, and one head for multiple moral hats

doi:10.1017/S0140525X17000632, e38

Bree Beal and Philippe Rochat

Department of Psychology & Institute of Liberal Arts, Emory University, Atlanta, GA 30322.

bree.beal@emory.edu psypr@emory.edu

Abstract: We support John Doris's criticism of "reflectivism" but identify three shortcomings: (1) his neglect of humans' evolved predispositions and tendencies, (2) his failure to appreciate that identity and responsibility arise first from parsing our world ontologically, in a process we call "existential framing," and (3) a potentially alarming implication of his "dialogic" model of identity formation: if identity is negotiated across diverse social situations, why isn't dissociative identity disorder more common?

With his latest book, John Doris expands upon the "situationist" interpretation he first outlined in *Lack of Character* (Doris 2002), now constructing a skeptical argument against a philosophical dogma he calls "reflectivism," the idea that "human agency consists in judgment and behavior ordered by self-conscious, accurate reflection about what to think and do" (*Talking to Our Selves*, Doris 2015b, Ch. 2, p. 1). In his view, evidence against "reflectivism" gives rise to a well-founded skepticism regarding morally responsible agency: if human agency does not meet "reflectivist" criteria, does this mean we are not morally responsible for our actions? Resisting this skeptical hypothesis, Doris proposes a "dialogic" model, where moral agency is achieved via a consensual process of negotiation with others. We are morally responsible agents, in Doris's view, insofar as our actions express our values (Ch. 7). Because values are determined through dialog with others and expressed through collaborative thinking and acting, people do have a kind of agency, and we can hold on to a socially distributed and consensual version of moral responsibility and agency by abandoning reflectivism in favor of dialogism.

We applaud Doris's integration of social psychology evidence into a moral philosophy debate, and we appreciate the desire to mitigate the consequences of scientifically induced skepticism via the proposed dialogic model. Moreover, we second his critique of "reflectivism." Anyone who holds to such a reductive view of agency should be challenged by the scientific literature suggesting that, for the most part, humans don't act as self-conscious, accurately reflective agents. Nevertheless, we feel that Doris's solution to this problem – his "dialogic" model of morally responsible agency – has several shortcomings. First, it fails to account for how the moral situation is constructed and constrained by humans' evolved predispositions and tendencies. Second, while focusing on social negotiation of values, Doris ignores how identity and responsibility arise implicitly as we evaluatively parse our world, in a process we call "existential framing." Finally, Doris does not address a potentially alarming implication of his situationist and dialogic model of identity construction: If identity is determined through negotiation across diverse social situations, shouldn't everyone suffer from something like dissociative identity disorder?

1. Born evaluators. To begin our critique with what is most obvious, Doris largely ignores the growing literature on the evolution and development of moral psychology, which would contextualize his account of the social negotiation of morally responsible agency. What do humans bring by way of evolved predispositions and constraints? Work from Jonathan Haidt, Frans De Waal, and others can help us appreciate the psychological significance both of our behavioral homologies and analogies with other species and of those features that are uniquely human. Any theory of moral psychology must be consistent with what we know of our evolved tendencies and developmental constraints.

In this connection, developmental evidence is especially revealing. Long before they can explicitly negotiate values with others, infants perceive and understand social dominance (e.g., Gazes et al. 2015; Mascaro & Csibra 2012; Thomsen et al. 2011), also preferring agents who express generosity over those demonstrating stinginess and unfairness (Hamlin et al. 2010). Finally, "inequity aversion" emerges reliably in children between the age of three and five years, in highly contrasted societal, religious, and economic environments (Blake et al. 2015; Rochat et al. 2009). Some asynchronies exist, but the developmental trend is universal and impervious to the drastically different ways children dialog about – and thereby co-create – values, across diverse cultures. Thus, in our view, it would be misleading to propose that prosocial preferences and inequity aversion are values created through social negotiation, while ignoring the contributions of our biology.

These examples illustrate an important point. If equity and prosociality are among the "values" toward which humans are predisposed, Dorisian agency is determined by something that precedes social negotiation. The emphasis of our critique, however, is on something even more basic than values: valuation itself. Humans are born evaluators, showing differential attraction and repulsion to objects in the world from birth. Even in the womb, fetuses develop preferences for their mother's voice or the smell of their amniotic fluid. We don't arrive through dialog at a value for our mothers; we are predisposed to seek comfort and sustenance, with corresponding feelings of warmth and connectedness, which babies evince from early on (e.g., Bigelow & Rochat 2006). Similarly, we don't have to be taught to enjoy sweet foods and (initially, at least) to have an aversion to very sour or bitter things (e.g., Rosenstein & Oster 1988). Through such innately specified preferences, we begin to make distinctions among foods, people, and other things, as we implicitly carve our world into qualitatively specified ontological categories. Thus, before we begin to negotiate explicit values, we already parse things and events in the world evaluatively. And this is crucial because, as we will argue in the next section, the value-laden specification of *what* something is, often carries implicit consequences for how we *ought* to treat it. The explicit negotiation of moral responsibility is thus typically a re*negotiation* of normative stances we have already adopted. To start with negotiation is to start too late.

2. Existential framing. In a 1963 interview, James Baldwin distilled the problem of race in America by attributing to white people a perverse "need": "What white people have to do, is try and find out in their own hearts why it was necessary to have a nigger in the first place, because I'm not a nigger, I'm a man, but if you think I'm a nigger, it means you need it."[1] What is this "need," and how does the invention of a new ontological category – a "n**ger" – fulfill it? We can start by simply recognizing that one might feel many obligations toward a "man," but far fewer toward a "n**ger." And before asking why this is so, we should first recognize that inventing a new ontological category was and is the most straightforward way to justify inhumane treatment of others, even as whites trumpeted values like life, liberty, and the pursuit of happiness. Typically, we don't spend much time arguing about such values. Instead we argue directly about identities and truths (e.g., What is the case? Isn't Baldwin a man?).

Consider, for example, the fraught issue of abortion. Both parties to the debate expressly agree on the values of human life, freedom of choice, basic care and protection for women and babies, and so on. Thus – moral grandstanding aside – the debate does not hinge on arguments over such values but on ontological claims that carry normative implications. What is a fetus? Is it an autonomous human being or part of a woman's body? Or something ontologically in-between? When and how does a fetus transition from one ontological category to another? Is abortion safe or unsafe? How painful are the various alternatives for mothers and fetuses? What is the burden of unwanted births on mothers, children, and society? And so on. Again, this is generalizable. We typically accomplish little by arguing that freedom is good, or that human life is precious – these are platitudes. Instead, most of the action is in the fight over ontologies.

How and why does this happen? Well before humans form explicit values, we are involved in relations of attraction/repulsion, intimacy, ownership, belonging/exclusion, cooperation/competition, etc. – relations that help us navigate our world successfully. For instance, the simple, preconceptual experience of trust and intimacy between people already imposes normative expectations for how such "friends" ought to treat each other – expectations that subsequently undergo social (re)negotiation. The same implicit formation of normative expectations occurs across

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

diverse contexts, as we form relationships with our families, homes, pets, or co-workers; as well as with those we perceive as disgusting or threatening. We and these entities are reciprocally co-defined in terms of perceived value, and these ontological determinations carry implications for how we ought to behave. Thus, before we ever negotiate "values," we already develop our identity by discriminating among people, places, beliefs, and things – and we do so in terms that carry normative implications. The adoption of shared values is a secondary abstraction and renegotiation of these implicitly formed normative expectations.

We propose to call this primary process "existential framing," a term emphasizing that our meaningful relationships with things in the world – in the Heideggerian sense of "existence" – shape or "frame" moral perception and judgment in all contexts. Existential framing has normative consequences that needn't be mediated by explicit value-negotiation. We simply treat "friends" one way and "foes" another; "pets" one way and "pests" another; "home" (my home) one way and "property" (a house) another. We only need to negotiate about explicit values in situations where these norms of behavior are contested.

Like Doris, we acknowledge the profound importance of dialog for establishing values – along with the pervasive dialogicity of human thinking. However, we feel that to indulge this argument over reflectivism versus dialogism is to focus on relatively superficial features, missing what might be more decisive for morality, agency, and identity. For example, Doris's suggestion that suffrage and civil rights are about groups' demand for "an identity that better expresses their values" (Ch. 6, p. 28) seems to miss the point. One could instead argue that the essential demand of suffrage and civil rights is not a claim about values (which values are specific to a gender or race?), but instead a claim *to value itself* – a repudiation of devalued and distorted identities and an assertion of reality. Or take the final chapter of *Talking to Our Selves*, where Doris summarizes the story of "Ishi," the last of the Yahi people in California. Doris attributes Ishi's loss of identity to "cultural devastation" (Ch. 8, p. 1). In so doing, he doesn't consider what it must have meant for Ishi's identity when, after having lived for years alone, he was forced by starvation to leave his home in the Sierra Nevada wilds and move into an anthropology museum. Ishi's loss was of course social and cultural, but it was surely more than this – it must have also been a loss of place, a loss of relations to forests, canyons, mountains, rivers, animals, and plants; a loss of relations of stewardship over and loving intimacy with these things; a loss of self-direction and self-determination. Damage to Ishi's identity must have been precipitated by more than a loss of culture or social ties because relations of intimacy, ownership, attraction, and stewardship with respect to non-human things also contribute to who and what we are. And such qualitative relationships also orient our agentive activity within a meaningful context. Doris's dialogic and "emphatically social" (Ch. 8, p. 1) model of morally responsible agency eludes this deeper dimension of existential framing, which we think is primordial in the determination of moral identity and responsibility.

3. Multiple moral hats. We conclude with a final provocation. Humans must negotiate a variety of social spheres, fulfilling roles that entail diverse responsibilities and parochialisms. Watching the news and reading the paper, we observe that being a loving father in the family sphere does not necessarily prevent one from being unmasked as a ruthless criminal in another sphere. In view of Doris's model of social and dialogic identity construction, this suggests to us that humans should end up developing a multiplicity of imperfectly aligned identities, corresponding to our roles in the distinct social situations in which we are embedded: submissive and obedient in one, domineering and violent in another. Nevertheless, even very great hypocrites tend to view themselves as having a single, coherent identity (and here, dissociative identity disorder is an exception that proves the rule). Our question is: How do we somehow manage to maintain a sense of coherence in our negotiated moral identity despite our constant switching of roles and moral "hats"? This crucial psychological conundrum tends to be neglected in moral philosophy, and continues to be so, even as Doris's situationist and dialogical account would seem to magnify the paradox of our inescapable moral ambiguities across social spheres.

NOTE

[1] See the film, *I Am Not Your Negro* (Peck 2016).

The participatory dimension of individual responsibility

doi:10.1017/S0140525X17000644, e39

Sofia Bonicalzi[a,b] and Mattia Gallotti[c]

[a]School of Advanced Study, University of London, Senate House, London WC1E 7HU, UK; [b]Faculty of Philosophy, Philosophy of Science and the Study of Religion, Ludwig-Maximilian University Munich, 80539 Munich, German; [c]Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK.

sofia.bonicalzi@gmail.com Mattia.Gallotti@gmail.com
http://www.mattiagallotti.com/

Abstract: Collaborativism is the view that moral reasoning is better when it is socially embedded. We propose that, when people take part in dialogic exchanges, they align in ways that open up novel avenues for sharing values and rationality criteria and, therefore, for exercising responsible agency. The hypothesis that collaborative interactions unfold through the alignment of minds and bodies helps articulate Doris's participatory approach to morality.

Scepticism about moral responsibility is the view that individuals are unable to fulfil a variety of prerequisites for responsible agency. A number of responsibility theorists have claimed that one of these prerequisites is accurate self-reflection. The difficulty with this view, as Doris suggests by drawing on research in cognitive and social psychology, is that seemingly rational choices can be determined by automatic processes. Because these processes are often accompanied by little cognitive elaboration and by inaccurate awareness of the causal history of behaviour, one might be tempted to deny the possibility of responsible agency. Doris, however, does not. He endorses the reflectivist's view that theory of agency provides the appropriate standpoint for assessing responsibility. Yet he argues that what we need for responsible agency is not self-reflection, which falls prey to agency defeaters and the practice of confabulation, but self-direction.

Self-direction is the capacity to express one's values with a justificatory status and a determinative role to play in practical planning (compare Bratman 2007). It is through *collaboration* that people's underlying desires for acting the way they do are articulated and expressed in the form of values. People come up with morally relevant explanations ("rationalisations") of their behaviour as part of a dialogical process of negotiation with other agents, a form of social discourse by which they give and ask others for reasons. This proposal echoes Scanlon's well-known conception that morality is a system of co-deliberation, motivated by the need to justify people's behaviour mutually (Scanlon 1998). Participation in dialogical endeavours provides a scaffolding for moral reasoning, but it may not take us any closer to knowledge of the true causes of behaviour than solitary thinking does. Collaboration has a practical, rather than an epistemic, role to play in moral reasoning: people forge their present identity and shape future behaviours by committing themselves to the values emerging from social interactions. Doris's argument for collaborativism thus puts us in a different position to counter the sceptic than canonical arguments for responsibility.

Collaborativism forges an interesting link between current trends in the philosophy of mind and morality and cognate fields of research, where the participatory dimension of individual thought and agency has become prominent in theorising about the

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

social mind (Kiverstein 2016). But the merits of this proposal can be strengthened by more fully articulating the explanatory potential of collaborativism and by suggesting a possible line of development. Doris's view is based on the belief that responsible behaviour cannot just be the expression of alleged stable traits of the individual, and that external factors are not confined to the role of mere distractors, as Real Self theorists have claimed (cf. Arpaly & Schroeder 1999). According to Doris, individuals are unlikely to have all of the internal resources they need to direct behaviour towards certain goals. The example of nicotine is a good case in point (Doris 2015b, p. 123). It would be hard to misconceive the positive effects of public health campaigns on changing individuals' attitudes towards tobacco consumption as the mere expression of one's own *inner citadel*. Responsible agency is tailored to acts of self-direction that facilitate the expression of values through collaboration. Put differently, moral agency is inherently collaborative. But what is the scope of this core claim of collaborativism?

Collaborativism is the view, not just that (moral) reasoning is socially embedded, but that it gets better *when* it is socially embedded (Doris 2015b, p. 119). Doris hints at two ways of interpreting this claim, which depend on how one takes sociality to facilitate the exchange of individual rationalisations. One is to say that dialogical endeavours are shaped by transfer of information between individuals. But there is more to the fact that optimal reasoning is socially embedded than the informational role of sociality, because mental activities recruit and process socially acquired information no matter whether the relevant activity is done socially or alone. Another interpretation would construe Doris's view as a version of the oft-quoted saying "two heads are better than one." People join forces because they can achieve better results as a group. Yet the fact that there are so many things in daily life that can only be done together does not mean that groups are smarter than individuals. Indeed, there is no conclusive evidence that group reasoning per se is more accurate, or delivers better results, than an aggregate of individuals working in isolation (Doris 2015b, p. 117). What makes moral reasoning optimal through collaboration is its explicitly interactive dimension. It is not the number of agents that shapes role-appropriate behaviour, but the fact that they do it *together*. The actions by which we hold someone responsible reflect values grounded in intersubjective exchanges in accord with the expectations and the norms that govern social roles.

Still, we may wonder: What is so special about doing things together, that is, creating narratives intersubjectively? While joining forces may not make us more likely to find out what the accurate drivers of behaviour are, thinking and acting together can allow us to create a shared space of thought and action, a mental space of communal access where the results of our negotiations become meaningful to us all and enacted accordingly. There is an important element of novelty here, which is hinted at by Doris without being fully elaborated. We contend that what is shared could simply not be reached for and attended to if not for our minds "meeting" in a suitable way. Cognising – or emoting, or intending, and so forth – does not necessarily enable optimal group performance, as we have seen, but it does give rise to new rationalisations, that is, mental contents which would be unavailable in thought, hence in dialog, to the single agents introspecting upon themselves. This is very much in line with Doris's remarks that there is a constructive element in building relationships of all sorts together, from the private up to the societal level, and that participation is the keyword in moral cognition and agency (Doris 2015b, p. 148). "Participation" is to be construed in the specific sense that taking part in the facts of life together does make a *difference* in that it gives us, each individually, new resources to live through those moments, to think and talk about, and act upon them.

These resources (values) are socially embedded and shaped in exchanges between individuals. More subtly, they become available by way of people sharing attitudes and dispositions and, therefore, creating a *shared* history of explanatory and justificatory narratives. How does this work? How can people move back and forth from their individual "confabulatory" perspective on things, to the collective system of morality that facilitates responsible thinking and acting? Recent work in social cognition reveals that when people engage in social interactions, collaboration unfolds through processes of alignment, whereby the relevant sharing is effected by exchanging and processing information at different levels (Dale et al. 2013; Gallotti et al. 2017). By adjusting to each other, gradually and dynamically, agents' repertoires of understandings, expectations and options for responsible action are expanded. Your reasons and criteria for holding someone responsible become part of my way of seeing things, and vice versa, so that *we* can operate on the basis of a collectively accepted infrastructure of values and rationality criteria.

The research on alignment and shared intentionality promises to be a valuable ally of Doris's collaborativism on several grounds. First, alignment does not necessarily require the recruitment of self-reflection. As interacting systems, individuals do not think and act together *as if* they were sharing perspectives on the world; they share perspectives in a pre-reflective manner. Of course, the relevant shared representations can rise to the level of conscious report, but they need not do so (Frith 2012). Second, while the emphasis in the social cognitive literature is often put on interactions being online and direct, shared perspectives can be reinforced by mechanisms of offline social cognition, whereby the community's moral norms are regarded as objectified and publicly available (Tomasello 2016). In line with Doris's suggestion, when people assume the perspective of their community, they experience both the pressure to offer sensible explanations of behaviour, as well as the sense of commitment essential for responsible agency. However, such collaboration can only be pursued by those who display the capacity to align at various levels of interaction. It is not by accident that we tend to exclude, from the set of subjects who are apt recipients of Strawsonian participant reactive attitudes, those creatures – e.g., psychopaths, non-human animals, and infants – who are (still) unable to engage in those dialogic interactions (Strawson 1962).

Arguing for a participatory dimension of individual responsibility, Doris provides an original response to the sceptic by sketching a theory of how morality works socially. Moral cognition and agency develop and acquire their contents in virtue of individuals' engagement in social interactions, by negotiating values and redefining the contours of their own personal biographies. We have suggested that more needs to be said about the mechanics of collaborativism, how it comes about, and why it helps articulate a cognitively oriented participatory approach to morality.

ACKNOWLEDGMENT
Many thanks to Gunnar Björnsson and Bryce Huebner for their generous comments.

Seeing for ourselves: Insights into the development of moral behaviour from models of visual perception and misperception

doi:10.1017/S0140525X17000656, e40

Daniel Collerton and Elaine Perry
Department of Psychology, Bensham Hospital, Gateshead, NE8 4YL, UK.
daniel.collerton@ncl.ac.uk elaine.perry@ncl.ac.uk

Abstract: Parallels from visual processing support Doris's cognitive architecture underlying moral agency. Unconscious visual processes change with conscious reflection. The sparse and partial representations of vision, its illusions, and hallucinations echo biases in moral reasoning and behaviour. Traditionally, unconscious moral processes are developed

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

by teaching and reflection. Modern neuroscience could bypass reflection and directly influence unconscious processes, creating new dangers.

Understanding how we distinguish right from wrong in our actions (moral agency) has suggestive parallels with deciding if what we see is real or unreal. In common with Doris's distinction between conscious moral awareness and the actual reasons that may underlie behaviour, there is a distinction within visual processing between what is subjectively seen in the mind's eye and the non-accessible visual processing that leads to that experienced image (e.g., Kanwisher 2010).

In this commentary we will use that commonality to explore three aspects of Doris's proposal: (1) Can conscious reflection alter unconscious processes? (2) What is the nature of moral representations in the brain and how are representations developed? (3) Can unconscious processes be the basis for free moral choices? Finally, we will consider some implications of brain-based morality.

It is currently believed that conscious visual perception derives from a sparse, internally generated, goal-directed model that predicts and is then constrained by input from the eyes (Collerton et al. 2005). The implication is that all perceptions, real or unreal, are fundamentally the same. In most cases images are constrained to be consistent with external reality – veridical – but on occasion the model goes wrong and things are seen that are not out there – hallucinations (Collerton et al. 2005). Individual reactions to hallucinatory experiences vary greatly, from a convinced belief that the experience is veridical, to an insightful awareness that the thing is not there, with many shades of understanding in between (Mocellin et al. 2006).

Deciding whether a specific perception is veridical or not will depend on a self-aware reflection which is usually developed in conjunction with other people – collaborative rationalisation, as Doris puts it. Doris (2015b) uses the example of confabulation in Ch. 4 to explore failures of reflective agency. The inverse occurs when people reflect upon hallucinations in a social context. For example, "rationalist" psychological treatments for distressing visual hallucinations emphasise a joint therapist-client investigation of the reality of experiences and the attributions attached to them, as the main avenue for improvement (Wilson et al. 2015). Significantly, in some cases insight alone may lead to hallucinations ceasing (Thomson et al. 2017), suggesting a sustained change in inaccessible internal representations as a result of reflection. Though this formal evidence comes from clinical disorders in which hallucinations are sufficiently common that they can be systematically investigated, anecdotal reports suggest that equivalent processes occur when people consider the non-pathological hallucinatory experiences that occur in some dreams, on the borders of sleep and wakefulness, or following sensory or sleep deprivation. As a contrast, "spiritual" (e.g., shamanic) accounts see these visions as reflecting a true, if alternative, reality (Luhrmann 2011), and may lead people to seek to actively encourage them. The individual attributions made of these perceptions, even when they have the same content, lead to great variation in how well people function in their lives (Waters et al. 2014).

Thus, in answer to the first question we posed, it does seem that conscious reflection in vision can have significant lasting effects on unconscious processes. By extension, Doris's fundamental architecture of non-conscious processes that bear a structured relationship to the environment and have outputs that are amenable to change following reflection both within individuals and with others, is plausible. Hence, moral behaviour may stem from a combination of non-accessible processes – the tendency to care, or protect, for example – and more reflective processes – how best to care or protect in a specific situation.

Neuroscience models of moral behaviour highlight the same distinction between unconscious decisions dependent upon orbital and ventro-medial prefrontal cortex, and conscious, reflective, moderation of these decisions, which is more reliant on dorsolateral cortex (Fumagalli & Priori 2012; Funk & Gazzaniga 2009; Mendez 2009). Furthermore, manipulations of neurotransmitters can have specific effects on moral behaviours. Dopaminergic agonists may make people appear to have more selfish behaviours, while serotonin boosting drugs may make people more caring (Crockett et al. 2015). However, while there is some consensus on the circuitry underlying moral judgements and behaviour, the nature of the representations and processes of morality within the brain is less established.

Does the analogy with vision therefore help us with our second area, the nature of the internal representations and processes which lead to moral agency? Visual processing is surprisingly sparse and partial given our subjective experience of a detailed consistent visual world. However, it is functionally good enough, despite being limited and inconsistent in some circumstances. For example, the phenomenon of change blindness, wherein most people miss major changes in the visual environment because they are not looking for them (Simons & Rensink 2005), illustrates that internal models do not have to be comprehensive or detailed in order to work well enough in everyday life. Visual illusions show that vision can get things wrong in some circumstances but again, still work well enough most of the time (Rees 2014).

Doris reviews the many biases in peoples' moral reasoning and behaviour. The analogy with vision suggests that the moral representations that people hold do not need to be consistent or detailed to still function well enough for them and society to get by. Thus, intellectual attempts to create a consistent moral framework may be at odds with how morality is represented within the brain. People may not follow a consistent moral framework, just as they do not see exactly correctly all of what is there, but they can still do well enough. Pragmatism may be the core approach of the brain to moral questions.

Doris makes the point that one can act in habitually moral or immoral ways. This would have the implication that the core of morality is not whether a specific action is moral or not, but whether the internal model that guides action is fundamentally moral or immoral. As with vision, there will be exceptions to the model (immoral hallucinations, if we could use the term), but the very fact that these can be reflected upon by the person, perhaps with others, and then recognised as deviant, supports the duality of processes that Doris proposes.

This line of thought leads us to consider how an internal moral model may come about. How do we learn to tell right from wrong? Evidence from developmental studies of vision suggests that though there are innate biases, much of perception is learnt by a combination of conscious and non-conscious processes in combination with other actors who act as instructors (Cohen & Salapatek 2013). This then places an emphasis on the potential benefit of moral instruction, and has echoes with traditional religious teaching on the development of an internal moral guide: conscience.

However, even granted that non-conscious aspects of morality can be taught and learnt, that does not necessarily mean that once a moral representation is formed, a person can still retain moral agency. In answer, it is striking that traditions that stress moral agency as diverse as Roman Catholicism and Buddhism both stress that moral behaviour comes from a combination of intuitive and reflective practices. The Catechism of the Roman Catholic Church (2000) states that "Moral conscience, present at the heart of the person, enjoins him at the appropriate moment to do good and to avoid evil" and that "Conscience is a judgment of reason whereby the human person recognizes the moral quality of a concrete act." Likewise, Buddhism "insists very strongly that there is right action, but that we cannot decide what right action is only by thinking about it . . . So that decision is not a decision in our mind alone, but an intuitive decision in our whole body and mind" (Luetchford 2001).

As Doris points out, the development of morality is a social process. Again from the Catholic Catechism, "The education of the conscience is a lifelong task. From the earliest years, it

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

awakens the child to the knowledge and practice of the interior law recognized by conscience." In Buddhism, [learning right from wrong] "is a very difficult task, and so it takes most of us, most of our lives to learn" (Luetchford 2001).

Analogies with vision therefore lead us to support Doris's general hypothesis. Moral frameworks may be incomplete and contradictory, but still functional. They are learnt to some degree, and though inaccessible to direct introspection, their outputs can be recognised and reflected upon, leading to change.

In our area of enquiry, psychological and biological studies arising from new models of hallucinations have taken an experience, once thought to be primarily a sign of madness, into the realm of normal visual perception. Conversations between informed philosophers, psychologists, and neuroscientists can work towards viable models of morality that incorporate bottom-up genetics, development, and biology; top-down social and educational influences; and the interplay of conscious and non-conscious processes. Might there be more 'common sense' if more such understandings can be relayed by responsible press and media?

The cognitive neuroscience of moral decision making is incorporating novel brain-based concepts, just as the cognitive neuroscience of visual perception did and does. Will the numerous individual variations in, for example, brain connectivity, the ability to suppress background information, and blood hormonal markers, mean that we can better understand the extraordinary variations in morality we encounter?

This raises the prospect of changing morality by modulating the brain directly, bypassing reflection. Thus, "Neuroscientists are now discovering how hormones and brain chemicals shape social behavior, opening potential avenues for pharmacological manipulation of ethical values" (Siegel & Crockett 2013) or "we aim to gather knowledge of the potential of tDCS [transcranial direct current stimulation] to modulate social functioning and social decision making in healthy humans, and to inspire future research investigations" (Sellaro et al. 2016).

But if morality can be directly changed without reflection and discussion, this raises a final question: Who will moralise the moralisers?

The dark side of dialog

doi:10.1017/S0140525X17000668, e41

Justin J. Couchman,[a] Gwenievere A. Birster,[b] and Mariana V. C. Coutinho[c]

[a]*Psychology Department, Albright College, Reading, PA 19604;* [b]*McLean Hospital, Harvard Medical School, Belmont, MA 02478;* [c]*College of Natural and Health Sciences, Zayed University, Abu Dhabi, United Arab Emirates.*
jcouchman@albright.edu mariana.coutinho@zu.ac.ae
gbirster@partners.org

Abstract: We agree that the self is constructed through a collaborative dialog. But hostile interlocutors could use various cognitive techniques to hijack the dialog, resulting in beliefs, values, and even selves that are out of line with reality. The implications of this problem are dire, but we suggest that increased metacognitive awareness could help guide this process to a truthful conclusion.

Doris builds his framework largely on the foundation of social psychology. He generally does not address the cognitive underpinnings of social psychology, on which the collaborative dialog is built and through which it can easily be exploited. Several well-established cognitive principles enhance and extend his idea, but are also notoriously used by advertisers, politicians, and other hostile actors. His framework is largely correct, but it is wrong to assume that everyone in the "dialog of self" is playing nice. In fact, many interlocutors are using cognitive principles to shape narratives and change values (and even selves) in surreptitious ways that may not result in the clarity that he envisions. The dialog itself will not naturally favor truth over falsehood, but it can be saved.

We would propose that the highest ethical standard in his system ought to be the process of increasing metacognition – the ability to self-regulate (to beat defeaters) and to avoid biases and hostile narratives (or meta-defeaters, if you will). Because metacognition is a skill and a tool, it is insufficient to simply say that "when defeaters obtain, the exercise of agency does not obtain" (p. 10). It is through metacognition that one can become better at recognizing and defeating defeaters. And it is doubly insufficient because by intentionally playing on human decision-making biases (e.g., heuristics, selective attention, hindsight bias, the first-instinct fallacy, etc.), hostile interlocutors can introduce a whole class of meta-defeaters that are engineered to surreptitiously change narratives, and thus values, and thus selves.

For example, suppose Candidate Creepy reads Simons and Chabris (1999) and learns about selective attention. He would discover that while focused intently on one event, humans ignore or do not notice things that would normally be salient. (In the original experiment, many participants counting the number of basketball passes in a busy environment did not notice a man in a gorilla suit walk into the middle of the scene, beat his chest, and walk off.) Thus, when journalists publish an in-depth investigative article about his indebtedness to foreign entities and business conflicts around the world, the Candidate might falsely but loudly declare that millions of people voted illegally in the recent election. Now the "collaborative" conversation focuses entirely on whether that claim is true, how he could possibly believe that, how others might believe it, etc. Temporarily interesting but ultimately meaningless content has monopolized the conversation. This will be internalized and go into forming the beliefs and values and selves of society, not the objectively more important and true information.

Suppose the Candidate claims he will build a giant dome over the country to protect us from aliens, and that the aliens will pay for it. The conversation ought to be about the merits of a dome, but instead we naturally anchor (Tversky & Kahneman 1974) to the outrageous idea of aliens paying for our dome, and by comparison the actual building of the dome seems relatively sane and doable. Also, the only way to selectively attend to the payment discussion forces us to accept, as a premise, that the dome will be built. Even attempting to move the collaborative negotiation (p. 35) back to the dome itself is impossible because the main objection – the cost – is subverted by the baseless claim that it will be covered by aliens. You can't talk about the cost without including that aliens will pay for it, and you can't talk about aliens paying for it without accepting that it will be built. Again, the narrative is hijacked and the beliefs, values, and selves that will form out of it will be based on misinformation.

Of course, dialog is even more vulnerable when we enter the fragile realm of human memory. Candidate Creepy might use hindsight bias (Fischhoff 2007), anecdotal evidence, confirmation bias (Wason 1960), or confuse correlation and causation. All of these effects just happen to play into the false claim that vaccines cause autism. How do you combat this falsehood? The natural reaction might be dialogic. Educate people to correct the myth! Unfortunately, Schwarz et al. (2007) describe how this leads to the backfire effect. Even if people are initially converted to the truth, many will end up continuing to accept the myth and will even believe it *more strongly* after several weeks have passed, not because of any surreptitious motives on their part, but because that is just how memory works. And the end result is that people believe vaccines cause autism, believe evidence supports this "fact," and form values and selves to match this belief. They might even go further and believe, again as core values and expressions of their selves, that scientists and journalists are misguided or evil for contradicting this "fact."

These three examples (there are hundreds more) might be called meta-defeaters, or processes by which defeaters are

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

created. They all prey on implicit processing, in which the people experiencing these effects would not be able to say to themselves "I am suffering from this psychological phenomenon engineered by Candidate Creepy." Instead they would believe, as a reflection of their values, that the Candidate was correct. But those beliefs were engineered explicitly and inserted into the dialog in order to change people into selves with values counter to the truth. Doris (sect. 2, para. 7) contends that we can rule out defeaters, and agency is exercised, if the action reflects the values of the actor. But what if your values are the result of meta-defeaters? Can a person be blamed for any of the "mistakes" above, given that each one would be a reflection of their unknowingly hijacked values?

This is not to say that Doris's dialog always leads to falsehood, but rather that it could equally lead to clarity or muddiness or even strengthened belief in falsehood unless some process is in place to guide it. Two brief examples will help illustrate this. Consider the primacy and recency effects (Ebbinghaus 1913), where people tend to remember the beginning and end of a dialog better than the middle. Managers use this effectively by providing positive feedback to employees at the beginning and end of a meeting and placing criticism in the middle. The employee will leave happy but will have a list of things to improve. Wonderful collaboration. More surreptitiously, advertisers place positive aspects of a drug at the beginning and end of a commercial, "hiding" the negative effects in the middle where they are more likely to be deemphasized or forgotten.

What is the difference between these two situations? Metacognition, also known as "thinking about thinking" or the ability to monitor mental states and use that information to control behaviors (Nelson & Narens 1990). Knowing that they will not remember the full content of any meeting – that is to say, being aware of their own memory limits – employees will take (or be given) written notes to aid memory. They'll still be subject to primacy and recency, but their preemptive information-preserving behavior will improve their understanding of the job. Commercial viewers will probably not take notes or be aware of their memory in that moment, and their understanding of the drug will actually get worse. Dialog plus a metacognitive component (recognizing your own mental abilities and incorporating that information into decisions) leads to clarity and a constructive self. Without metacognition there is nothing to protect against ignorance, misinformation, and values that are out of line with reality.

Note that this is not just a passive phenomenon, but a skill that can be improved. This is why we suggest that it is insufficient to merely accept defeaters; they must be actively defended against. Consider one final example of a cognitive bias called the first-instinct fallacy, which is the false belief that first instincts are special or more likely to be correct. It is well known that taking an exam is a form of dialog that can be constructive and lead to better memory/learning/clarity (Roediger & Karpicke 2006). But many people have learned, and even been told by teachers, that their first instincts ought to be trusted. Despite overwhelming research to the contrary, the belief persists (Kruger et al. 2005). If someone trusted their first instinct and got a question wrong, Doris (sect. 2, para. 7) would likely contend that defeaters obtained and the act was not agentic. But clearly this is insufficient, because there is a method to overcome the defeater and the person would be held responsible for their choice whether they overcame the defeater or not.

Couchman et al. (2016) used a simple method of having exam-takers keep track of their level of confidence in each answer, specifically to guard against the transience of memory and several of Tversky and Kahneman's (1974) heuristics. Confidence ratings are a common form of metacognitive assessment, and they found that these in-the-moment self-assessments were predictive of objective accuracy. In fact, when deciding whether to change an answer or not, confidence ratings were a better guide than first instincts. Contrastingly, assessments made before and after the exam – explicit judgments subject to the fragility of memory and the cognitive biases described above – were less accurate. Overall, they found that defeaters could be overcome, misconceptions avoided and biases circumvented, by tracking confidence to increase the availability of metacognitive information and aid the decision-making process.

Thus, implicit self-monitoring, a reflective process in humans and even some animals (Smith et al. 2012) that uses working memory and can be trained (Coutinho et al. 2015), can correct explicit biases that can result from dialogic collaboration. The same process could extend to any "collaboration," to ensure that values are represented and formed in ways that are not based on known errors (even if they are unnoticed defeaters).

Finally, we would note that metacognition is closely related to self-agency (Couchman 2012), self-awareness, and importantly to theory of mind, the process of understanding others' thoughts and taking their perspective (Carruthers 2009; Couchman et al. 2009). Metacognition is the primary tool we use to overcome self-ignorance. As Doris (sect. 9, para. 5) states, "agency-impairing self-ignorance is considerable, but people are collectively possessed of epistemic assets fit to ameliorate it." This is partially correct, but without working to improve our uncertainty (or ignorance) via our monitoring abilities, we would have no way to judge when defeaters have obtained and no way to know which collaborations increase agency and which might ameliorate it. Similarly, theory of mind is the primary tool through which meta-defeaters and surreptitious dialogic tactics are generated. By taking the perspective of others, and understanding well-established cognitive techniques, a hostile interlocutor can change beliefs, values, and even selves to be out of line with the truth.

To use a metaphor: Doris essentially describes how journalism (collaborative dialog) could build a great society (the self). We fully agree, but at the same time we welcome him to the war. Propaganda (via meta-defeaters) is already being used to harm the great society, through the very channels he proposes will save it. There is no check whatsoever to ensure that dialog alone will lead to truth, and in fact we know it can easily be used to create selves with values out of line with reality. Thus, we are compelled to improve our fact-checking and editing process (metacognition) to regulate journalism and make sure it steers us toward a clarifying truthful dialog.

The self is indeed a function of collaborative rationalization, and Doris (sect. 8, para. 7) correctly points out that agency (and by extension the self) does not require freedom from influence – an impossible criterion – but rather mutual influence. But with collaboration comes conflict, and it is a grave mistake not to recognize this. Collaboration could lead to Utopia, but hostile actors could also steer us into the Wild Wild West, or worse. And as such, the highest ethical standard ought to be to increase metacognition. To increase self-monitoring and fend off untruthful narratives that intentionally play on human cognitive biases. This holds true in all human activities, even in extreme cases like addiction (Hajloo et al. 2014) that Doris discusses.

Moral agency among the ruins

doi:10.1017/S0140525X17001157, e42

David Dunning

Department of Psychology, University of Michigan, Ann Arbor, MI 48109.
ddunning@umich.edu **https://sites.lsa.umich.edu/sasi/**

Abstract: Doris suggests thought-provoking directions for rehabilitating moral agency within a self that is unaware and incoherent. These directions suggest more radical proposals. First, moral reasoning may serve many different functions beyond merely expressing a person's values. Second, social collaboration may not focus on moral reasoning as much as it does on the "defeaters" of that agency. Ultimately, moral

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

agency may not reside in the individual but in social communities or within external situations.

In his new book, John Doris takes on an intimidating interdisciplinary challenge: How does one retain a philosophically intelligible account of moral responsibility that is consistent with current themes in psychology and cognitive science? The contradictions are stark between the two. As Doris notes, typical philosophical accounts of the moral reasoner assert that people are fully functioning agents making moral choices after careful and conscious deliberation. In a phrase, they are reflective moral agents (e.g., Annas 1993; Brink 1992; Tiberius 2002) in full control of their moral choices, basing those choices on visible and diligent ethical reasoning, and fully self-aware of the basis for their moral decisions.

Against this logic, contemporary research in psychology, cognitive science, and neuroscience shatters the portrait of the self-necessary for such a "reflectivist" account of moral reasoning. Instead, extant research suggests that the self is anything but a coherent, reflective, and rational being (Stich 1990). To be sure, people experience themselves as conscious and deliberative organisms, but the real action leading to their moral choices lies elsewhere (e.g., Haidt 2001).

In particular, people's choices are importantly determined by System 1 processes, that is, mental operations that are quick and that often operate under the level of conscious awareness and control. These are distinct from System 2 processes, conscious and effortful cognitive operations that the moral reasoner can guide and govern (Kahneman 2011a; Sloman 1996). The operation of System 1 suggests that the true causes for people's moral behavior can be distinct from, or incongruent with what the moral reasoner believes them to be. System 1 can "defeat" moral agency so analytically constructed in System 2.

In the first half of the book Doris guides the reader through decades of empirical data that have accomplished this shattering of the agentic self. He takes readers through findings showing that people often produce mistaken accounts of themselves and the causes of their behavior, showing how conscious thought and judgment about the self are often filled with confabulation, illusion, and misunderstanding (Berlyne 1972; Dunning et al. 2004; Johansson et al. 2005; Nisbett & Wilson 1977). The argument Doris makes is rigorous and comprehensive, and contains, I believe, one of the most coherent and instructive narratives about what twentieth-century psychological science has to say about self and agency – in both senses of self that William James talked about: the "I" as the doer and the "me" as the object of reflection (James 1890/1950).

That said, the second half of the book is where the more difficult segment of Doris's task begins: After all this shattering of the agentic and self-aware agent, can one rehabilitate the idea of moral agency and responsibility among the ruins that are left? Here, Doris is less successful simply because he is less complete. He offers directions to take rather than destinations to inhabit.

Those directions, however, are intriguing. The first is the proposal that moral decisions are first and foremost designed to be expressions of a person's values. The second is to reject moral agency as a completely individual exercise and instead emphasize "collaborativism" among people. People do not conduct their moral reasoning in isolation. Instead, they discuss and debate it with other people.

Let's evaluate each of these proposals in turn. The first is that moral choices are primarily expressions of a person's values. This is an intriguing idea, but Doris leaves it underdeveloped. First, why would System 1 processes lead to moral behavior favoring expression of values over all other possible alternatives? There are many to consider.

Dan Katz, in his classic article on attitudes, noted that attitudes and opinions, like all human responses, serve many different functions (Katz 1960). To be sure, people endorse certain attitudes because those attitudes express the values that they hold most dear (e.g., of liberty, individualism, or equality). But there were other functions as well. People might hold certain attitudes because they were the ones that tune the person best to objective truth. Attitudes can also protect and bolster the ego of the person, or assist that person to fit in as a proper member of the groups he or she wishes to affiliate with. These other functions seem just as plausible an engine for System 1 driven moral choices as does value expression. So why privilege value expression?

Second, in emphasizing value expression, Doris potentially makes a thought-provoking implicit proposal well worth following up: People are more concerned about the actions they choose than they are about the outcomes those actions might produce (Cushman 2013; Dunning & Fetchenhauer 2013). That is, people want to express being the right person by choosing the correct behavior rather than by ensuring the right outcome. I have seen such behavior in my own lab. People choose to trust complete strangers with their money even though on average they expect they will never see that money come back to them. They do not seem to be concerned with the outcome of losing the money as much as signifying, through their behavior, the "right" social attitude, namely that they respect the character of the stranger until proven otherwise (Dunning et al. 2014; 2016). As such, one can ask under Doris's scheme whether one can cross out not only reflection as an important component of moral behavior but also consequentialism.

But the more important move is the one toward collaborativism, the proposal that moral agency arises not from individual reasoning but thinking among people. Moral reasoning here becomes a social product, one shaped and guided by action and talk among people as they live their daily lives trying to get along in something better than a nasty, brutish, and short interplay.

Doris proposes that authentic moral agency emerges from this social interaction and discussion, that people achieve accurate and reflective self-awareness of the bases that inform their moral choices. First, he notes correctly that groups tend to produce more insightful and sound solutions to problems than do individuals (Hill 1982; Schwartz 1995). Second, he notes that confabulation, if critiqued and directed by others, may come closer to accurately citing the values that truly drive moral choice. As such, through collaborative social interaction, people come closer to moral reasoning that reflects the true causes of their moral behavior. They become more authentically self-aware.

These are reasonable and testable proposals, but there are other, potentially more radical proposals hiding in plain sight within recent psychological literature. For one, despite the move toward collaborativism, Doris still holds to a rather individualistic and Western vision of agency, one in which people freely choose without too much influence from outside forces. People still act as captains of their own decisions, thus imposing their will on the external world. In psychology, this is known as disjoint agency, the type of agency usually presumed in Western cultures (Savani et al. 2008; 2010). Other cultures hold to a more conjoint model of agency, in which people strive to harmonize their actions with outside forces and constraints, usually social ones, which they are surrounded by. As such, people still view themselves as agentic, but not just as complete free agents (Savani et al. 2008; 2010). It would seem to me that if people are collaborative, as Doris speculates, they would be more likely to shift to a conjoint stance in their moral agency than a disjoint one.

But more important, Doris emphasizes the impact of collaborativism on moral reasoning. I would instead speculate that collaborativism would have more of an impact on what people actually do (Henrich et al. 2010). The major impact of social collaboration would be on those System 1 defeaters that make people act in ways incongruent with their reflective moral agency. After all, a key insight of recent work on prosocial behavior is not that people go out of their way to be compassionate and generous. Instead, they are kind to others because they are bending to social rules and obligations.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

As such, prosocial behavior is often not so much about giving as it is about giving in – to norms and social roles (Cain et al. 2014). For example, if Salvation Army volunteers expressly ask for holiday donations outside of a supermarket, the number of people donating increases significantly. So, however, does the number of people exiting out some other door to avoid the volunteers (Andreoni et al., in press). Apparently, people will give, but it may not reflect their moral preferences. Our work on trust behavior also exemplifies this. People do not go out of their way to trust other people. Instead, the key factor is that they feel so anxious and tense about not trusting the other person. Trust is what they should do, not necessarily what they want to do, and so they give in (Dunning et al. 2014).

Here's another example about how collaboration may center more on agentic defeaters than agentic reasoning: If you ask people whether they will challenge someone making a sexist or racist comment, they typically claim that they would confront the person directly and immediately. However, if you give them an actual encounter with such a person, they act instead by ignoring or deflecting the comment (Kawakami et al. 2009; Swim & Hyers 1999; Woodzicka & LaFrance 2001). At the moment of moral truth, a lifetime of collaborative social training and interaction appears to have taught people to be polite and respectful of others rather than to stand up for their personal values (Brown & Levinson 1987; Goffman 1958; 1967). Politeness trumps (i.e., defeats) moral values. Of course, there are socially proscribed occasions when argument and contention are expected (cable television, anyone?), but those instances are few and well-defined.

What this may ultimately mean is that moral agency may not reside in individuals but rather in social groups that make up, teach, and enforce the social rules that people live by. Or instead, moral agency may arise in the cues of social situations that signal one value over its alternatives (Lindenberg 2012). People in economic games, for example, adopt rather different stances toward cooperation versus selfishness simply depending on the name of the game they are incidentally playing – for example, "the community game" versus "the Wall Street game" (Liberman et al. 2004). The simple choice architecture of the game itself can suggest a moral value. In simple dictator games people tend to share the money they have with other people. However, adding the option of taking the other player's money stops that sharing. Instead, the modal choice is for players to take the other player's money. Unprompted generosity turns to unbridled greed (List 2007).

In sum, *Talking to Our Selves* is a pleasing and thought-provoking tour of the philosophical and psychological issues surrounding agency, self-awareness, and moral choice. It is a delight for those wishing a well-curated tour of past scholarship on these issues, particularly psychological research. It also sets a fine table for a potentially roiling but useful discussion about these issues yet to come. I hope it provokes those discussions. It would be collaboration well worth having.

On properly characterizing moral agency

doi:10.1017/S0140525X1700067X, e43

Blaine J. Fowers, Austen R. Anderson, and
Samantha M. Lang

Department of Educational and Psychological Studies, University of Miami, Coral Gables, FL 33124.

bfowers@miami.edu **aanders8@yahoo.com**
samantha.lang718@gmail.com **www.blainefowers.com**

Abstract: Doris (2015b) develops a theory of moral agency to avoid a skeptical challenge arising from psychology studies indicating that (im) moral behavior is caused by trivial situational factors. His theory is flawed in attending only to situational influences on behavior and neglecting individual differences such as moral identity and virtue. A focus on individual differences in resilience to influence from trivial situational factors defangs the skeptical challenge and offers a better account of moral agency.

Doris invites scholars to investigate ethical questions "in light of the best ongoing scientific picture" (Doris 2015b, p. 12) to make their theories more empirically credible. As psychologists who work at the boundary of psychology and philosophy, we heartily endorse this goal. He exemplifies this through a philosophical discussion of numerous social psychological experiments, concluding that there is room for "skepticism about morally responsible agency" (p. 1) because there are numerous "defeaters" of agency (e.g., ballot order effects). He bases his problem diagnosis – that defeaters undermine morally responsible agency – on evidence indicating that apparently (im)moral actions are caused by minor environmental factors outside the actor's reflective attention. The worry is that individuals' reason and choice are "bypassed" by a cause that cannot provide a rational justification of the act. If Doris's diagnosis is correct, we may have a serious problem. In response, he defends moral agency as the expression of one's values, and champions collaborative reasoning as a key source of agency.

We suggest that Doris's worries about defeaters are overstated. Central to his overstatement is that the experimental effects that he cites are generally mild and inconsistent. Ballot order effects are one of his favorites, but the literature is far less worrisome than he suggests. The effects occur more in low information and low visibility elections (Pasek et al. 2014). Clearly, the fewer clear reasons to vote for a specific candidate, the more ballot order will sway voters.

Doris wisely recognizes and discusses the mild, aggregate effects problem: "To be sure, identifying statistically small effects does not allow confident conclusions about particular outcomes for particular individuals" (Doris 2015b, p. 63). Yet he still tends to overstatement. He concludes that "In sum . . . evidence of incongruence is readily obtained," which "make[s] plausible the supposition that incongruence is *widespread* in everyday life" (p. 61). It is unclear how he translates group differences in experiments with strangers into widespread occurrences in ordinary life. He makes this leap partly by adverting to a confabulated cumulation of small effects because "there could be *many* goofy influences in any particular instance . . . [and] the aggregate effect may be quite potent" (p. 64). This goes considerably beyond the evidence. There is precious little evidence for multiple situational influences operating simultaneously, and virtually none that multiple situational factors move individuals in the same behavioral direction. It is just as likely that situational factors cancel one another out as cumulate.

Because Doris builds his philosophical argument on social science data, it is important to apply scientific skepticism to philosophical skepticism. He states the skeptical challenge thus: "If there is general difficulty in ruling out defeaters, skepticism about agency ensues" (Doris 2015b, p. 65). This is a strong claim; too strong, given the available evidence. Given the aggregate effects and statistical likelihood of the influences, the most that can be said is that defeaters show that not all choices by all people are governed by justifiable moral reasons.

Doris's overstatement of this problem may be partly due to different forms of argument in philosophy and social science. Philosophers incline to search for absolute truth, which fuels the skeptic's challenge. Social scientists do not care much for absolute truth and speak in terms of tendencies and probabilities, and this is the appropriate language for the data Doris cites. In this language, there is a small probability that a given individual act will involve "bypassing" the actor's reason. Furthermore, this mild influence is typically observed in rather trivial circumstances with strangers. The skeptical argument has no purchase here because social scientists *assume* that humans are imperfect reasoners and that individuals vary in reasoning quality. This is not

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

as surprising or earth-shaking as Doris makes it out to be. Even if one reasonably grants that everyone has some vulnerability to defeaters, all that is being conceded is that humans are imperfect moral reasoners, which should not alarm us. This imperfection is only a blemish against agency if one imposes a perfectionistic requirement that choices are always conscious and well-justified. These are unreasonable assumptions. To create an empirically credible moral psychology, Doris must let go of such psychologically unrealistic starting points.

The skeptical challenge is further defanged when we consider what Doris entirely neglects: individual differences in moral agency and in susceptibility to defeaters. His singular focus on situational influences and his omission of the moral psychology of individual differences distorts his problem diagnosis. His choice is not surprising because this omission is pervasive in social psychology, the empirical foundation of his argument. What would happen if we took individual differences seriously in the experiments Doris cited? No one knows because they are rarely studied in this literature. Although social and personality psychologists agree in principle that behavior is best explained by a combination of situations, traits, and their interactions, actual person-X-environment studies remain the exception rather than the rule.

To clarify the importance of individual differences, suppose that the vulnerability (or resilience) to defeaters itself is treated as an individual difference variable. Thus, some people would be more likely to be influenced by defeaters (i.e., act based on trivial situational influences) and others would be less likely to be so influenced. That is, some individuals are better at maintaining and expressing their value commitments than others (a capacity that is central to Doris's viewpoint). The apparent crisis raised by agency skeptics turns out to be an unsurprising variation in capacity to maintain value commitments. In fact, a primary domain of moral psychology investigates this consistency in value commitment under the rubric of moral identity (e.g., Blasi 2005), which Doris also curiously fails to mention.

One interesting counterexample of individual difference neglect is a study that Doris co-authored, which he cites as an example of induced disgust leading to more punitive moral judgments (Cameron et al. 2013). In fact, there was *no main effect* for disgust induction in this study. There was an interaction of disgust and emotion differentiation such that participants low in emotion differentiation were affected by disgust, but those high in emotion differentiation were not. We do not know why Doris failed to accurately describe this individual differences result, but we do know that this oversight favors his hypothesis and indicates his disregard of individual differences.

Another illuminating example of this problem is Doris's discussion of Mischel and his colleagues' work on delay of gratification (DG) in children. Interestingly, Doris only cites Mischel's early experimental work, which focused on situational influences on children's DG (e.g., Mischel et al. 1972). This work fits Doris's thesis. However, he does not cite the later studies of individual differences in DG capacity, which show that DG and rejection sensitivity predict educational attainment, self-worth, interpersonal functioning, and lack of drug use over a twenty year period (e.g., Ayduk et al. 2000). This work suggests that agentic capacity is, at least in part, an individual difference variable, which unsettles Doris's thesis.

A simple thought experiment on ballot order effects can further illustrate the importance of individual differences in agency. We propose a thought experiment because we do not believe that anyone would think the outcome is sufficiently in doubt to recommend actual data collection. First, we stipulate a small ballot order effect, consistent with Doris's presentation. Second, let us examine a consequential, high visibility contest: the 2016 U.S. presidential election. It must be consequential and visible or Doris's worry about moral agency is moot. The less the election matters, and the less people know, the less a vote can be considered a question of moral agency. Third, we divide the voters into three groups: committed Trump voters, committed Clinton voters, and undecided voters. Fourth, we randomize the presented ballots, with half the ballots listing one candidate first and half listing the other candidate first. The outcomes seem obvious. The small, stipulated main effect for ballot order would almost certainly be strongly qualified by an interaction. The ballot order effect would be far stronger in the undecided group than in the committed groups, illustrating the central role that individual differences (in value commitment) play in agency. Those with clear, strongly held values will be far more defeater resilient than those with ill-defined or weakly held values.

It could be objected that this is a too-easy counterexample, but there is a general case available in virtue theory, which provides an explanation of resilience to defeaters of agency. On Aristotle's (340 BCE/1999) view, there are multiple overall character types, three of which can illustrate its predictions. Individuals with virtuous characters are highly resilient to defeaters because they have strong moral commitments and have made moral agency habitual. Continent characters know what they ought to do and generally act accordingly, but they are not as firmly committed to acting morally as the virtuous, which makes them somewhat more vulnerable to defeaters. Although incontinent characters know what they ought to do, they are weakly committed to acting morally, making them highly vulnerable to defeaters. Behavioral research on virtue and character is just getting under way (e.g., Lefevor & Fowers 2016; Meindl et al. 2013), but the expectation is that virtue and character will directly reduce defeaters' influence as well as moderate defeater effects. If so, this changes the picture substantially. Some people are very vulnerable to defeaters, some are somewhat vulnerable, and some are relatively invulnerable. Importantly, these individual differences in agentic resiliency do not rely on reflection because they manifest in both automatic and reflective action.

Doris could claim that he has already dispensed with virtue and character in his previous book (Doris 2002). He argued, based on social psychology experiments, that small situational factors influenced participants to act more or less morally (primarily whether they helped a stranger). He claimed that these effects mean that moral character cannot be very important. That argument has been thoroughly contested theoretically (e.g., Kristjánsson 2013) and drastically undermined empirically in a meta-analysis showing that the experimental effects he cited as evidence against character are empirically insufficient to rule out character as an explanation for helping behavior (Lefevor et al. 2017).

Just as the addition of individual differences modifies the problem diagnosis, it also undercuts the value of Doris's remedy. If differences in the capacity to maintain value commitments is part of the problem, then a clear conception of individual differences must be included in any moral psychology. Therefore, Doris's quick dismissal of virtue theory and utter neglect of moral identity theory are misguided. Both theories can account for the variations in vulnerability to defeaters of morally responsible agency. In addition, moral identity and virtue theories can also explain Doris's definition of agency as behavior that expresses the agent's values: The stronger one's moral identity or character, the more one will have well-defined values, and these values will be expressed with greater frequency and automaticity. This is a deeper and more comprehensive account of moral agency than Doris offers.

We do find Doris's argument for a collaborative understanding of reasoning and agency very congenial. However, his view on collaborative reasoning must also give appropriate weight to morally relevant individual characteristics. Clearly, some people will have greater capacity for collaborative reason than others, again highlighting moral character or identity. For us to engage in the best collaborative reasoning possible, we must cultivate the excellences of collaborative reasoning (e.g., openness and honesty). Moreover, we cannot ignore the aims of human collaboration. We agree with Doris that collaborative reason is invaluable to moral agency, but it can also be a source of moral blindness, as amply demonstrated by the "collaboration" that made the totalitarian slaughters of the

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

twentieth century possible. Although we can only raise these thorny issues here, we do not think it is possible to describe or explain moral agency without considering the character of the reasoners or what makes the ends toward which they reason worthwhile.

What does agency afford the self?

doi:10.1017/S0140525X17000681, e44

Bradley Franks[a] and Benjamin G. Voyer[b]

[a]*Department of Psychological and Behavioural Science (PBS), London School of Economics & Political Science, Queens House, 55/56 Lincoln's Inn Fields, London WC2A 3LJ, UK;* [b]*ESCP Europe, UK.*

b.franks@lse.ac.uk **bvoyer@escpeurope.eu**
http://www.lse.ac.uk/PBS/People/Dr-Bradley-Franks
www.benvoyer.com

Abstract: We welcome Doris's dual systems, social account of agency and self. However, we suggest that a level of affordances regarding agency is interpolated between those dual systems. We also suggest a need to consider joint ("we") agency in addition to individual ("I") agency, and we suggest a more fundamental role for culture in configuring both the values entering the dialogue that generates the sense of agency and self, and the nature of the dialogue itself.

Doris offers a thought-provoking and persuasive account of the interrelations of self, agency, and moral action. We especially welcome its emphasis on social, dialogical, and evolved origins of agency, which echo to important developments in the field (Voyer & Franks 2014), suggesting that the sense of self, in arising from agency, is an evolved and cultural construction (Franks 2014). Its broadly pluralistic approach to determining which aspects of context have a role in this process, and how, is congenial to an empirically grounded approach to agency and the self. However, we suggest that the account is incomplete in three significant ways.

First, Doris assumes dual systems – implicit and explicit in their representations and processes – that govern moral judgement, behaviour, and explanation. The former involve non-conscious, rapid, and relatively effort-free processes and judgements whose reasons are not open to introspective access. The latter involve conscious, slower, effortful processes and explanations that are framed in terms of narratives based on a person's values (in particular, as they are reflected in and debated through dialogical processes with those around them). This is, of course, a widely held view (e.g., Kahneman 2011b; Stanovich 2011). Doris does concede that this may be incomplete, that there may be representations and processes intermediate between these levels. We concur with the role for implicit and explicit systems, and following Voyer and Franks (2014), we suggest an intermediate level is interpolated between them.

To see this, it is important to note what Doris's account of agency does not attempt to do. Doris focuses on what we might call "revealed" agency, by analogy with revealed preferences in economics: agency as revealed through the values that a person expresses through the actions they perform. We agree that the fine-grain of agency and responsibility should be connected to actions under the relevant values. However, we suggest that such a behavioural characterisation misses part of the puzzle about agency, which is that it is also felt or experienced, processed, and perhaps represented as an agentic state *per se*.

We suggest that it is important to provide an account of such experiences of agency, because they are a large part of the psychology of agency, they figure significantly in people's own normative explanations and justifications, and connect directly to the sense of self. Agency is simultaneously first-person and second- and third-person: It is a first-person subjective experience (e.g., the feeling that "I did this") or a second-person experience (e.g., "we did this": see below), which is grasped in part via dialogue with third-person views of others (e.g., "people do this").

For the sense of agency, the sharp distinction between explicit and implicit processes is incomplete: Voyer and Franks (2014) suggest that three different modalities are involved in its assessment: implicit (a "feeling" of agency), intermediate (a "perception"), and explicit (a "judgment"). The intermediate level of self-*related* processing is in line with recent developments in the field of cultural neuroscience (Han & Humphreys 2016; Sasaki & Kim 2017). It may involve composing complex plans and actions from less complex ones but does not, in itself, require conscious awareness of that agency (e.g., holding a cup whilst someone else is pouring a drink in it). The three modalities are represented in different mental formats, each with their own characteristic form of intention. Cues about whether one is or should be viewed as the agent of an action are based on affordances (Dreyfus 1985; Gibson 1977; Franks 2011; Kitayama & Imada 2008), informational relations which may be based on evolved, perceptual, social, and cultural foundations. Affordances related to one modality may input to processing another, and the different modalities may generate mismatching conclusions about agency for a given action. Whereas for Doris the possible mismatch between levels is worrisome (suggesting defeaters for explanations for action), for our account it motivates a key form of dialogue between different cues for agency regarding an action.

This has two implications for Doris's view. The first is that the sense of agency – and its absence, e.g., institutionalisation, anaesthesia, extreme conspiracy theory belief (Franks et al. 2013; 2017) – is a situation-dependent state, which often takes characteristic forms in different social and cultural environments, so that in aggregation it can appear to be a stable trait related to particular ways of acting. It is important in explaining and justifying behaviour, though its normative evidential status is fraught. The second is that agency, as connected to subjective experience, is more various than Doris suggests.

This leads to our second point: Doris's social account of agency is not social enough. He focuses on social causes and consequences of agency, but does not take into consideration the social nature of its experience, which in turn can affect its expression. He considers "I" as a possible agent, and we concur with his broadly dialogical account of this. He does not, however, consider another important form of agency – "we" or joint/shared agency. People engage daily in many forms of joint actions and projects, ranging from walking together, sitting side by side on a commuter train, having dinner, playing football, etc. In these cases there is a plural subject, a sense or experience that "we" are the agent. Indeed many of the cases of moral decision making that form the focus of empirical studies on which Doris draws are, in their real-world manifestations, more properly thought of as cases of joint or shared moral decision making, for instance two surgeons operating on a patient and having to make a potentially life-threatening decision. Ultimately these are decisions that are made with the implicit or explicit belief that real or imagined others concur or demur.

There are complicated debates on how to understand joint agency, for example its relations to individual agency (e.g., Gilbert 1989; Searle 1995; Tuomela 1995). On the one hand, there is the possibility that joint agency reduces to combinations of individual agency ("we"="I am doing this, and so are you," that is, we are both performing individual actions simultaneously). On the other, joint agency might be irreducible, *sui generis* ("we" just="we are doing this," without the possibility to separate individual actions concurring to the desired outcome). The possibility of such different senses of "we" has a long history in social and cognitive psychology (e.g., Brewer & Gardner 1996; Swann & Buhrmester 2015; Tomasello 2009; Turner et al. 1987). Whether joint agency does reduce for all or some states, it is likely that these two forms of joint agency generate qualitatively different experiences of agency.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

There are two implications to note here. One is the relation to different moral sentiments in judging and cooperating with others. Bloom (2016) has recently taken aim at the widespread notion that empathy is intrinsically morally valuable, an inevitable precursor to prosociality and cooperation. Empathy, the sense of sharing another's experience (and sense of agency in relation to external events), he argues, can in fact diminish moral action. By contrast, sympathy, the sense of being aware of another's experience as different from one's own (and of their differential agency in relation to external events), may be more likely to generate moral action. When they issue in joint remedial action against the suffering of the other, we suggest that these different states lead to different morally valenced forms of agency. Empathy reflects irreducible "we": both have the same experiences, the same sense of agency and the same capacity and responsibility for remediation. Sympathy reflects reducible "we": the "I" and the "you" are separable and represented and experienced as separate, with separate but possibly interconnected capacity and responsibility for remediation. In sum, sympathy reflects our reducible perspective on shared agency ("we"="I am doing this, and so are you"), whilst empathy reflects our suggested irreducible perspective on shared agency ("we" just="we are doing this"). The second implication is that the interpretation of joint and individual agency, and their relative preponderance in explanations of behaviour, are themselves related to different cultural tendencies to view oneself as more or less interdependent or independent on others regarding action and moral judgement.

This leads to our third and final point. Doris proposes that a dialogue of values generates agency and the sense of self, and notes that this relates to the wider culture in which the person lives and acts such that the values that come to characterise the narrative of the self may differ from one culture to another. We concur, but suggest again that this underplays the significance of the social and the cultural context.

Culture plays a key role in the nature and functioning of the values, which Doris suggests are constitutive of agency. Values vary across cultures, and behaviours that are prescribed in one culture may be proscribed in another (e.g., Schwartz & Sagie 2000). The degree to which those values are experienced as constraining and directing, also varies: cultures differ in "tightness" and "looseness" (Gelfand 2012; Uz 2015). Hence, the degrees of freedom in the dialogues that define agency will also vary between cultures: For a given action, it may range from a top-down imposition of widespread and tightly adhered-to norms to an interplay of equals. This has implications for agency at all three modalities, including the narratives generated to ascribe or avoid responsibility in the context or morality judgements and attributions. Voyer and Franks (2014) suggest that the dialogue between individual and social and cultural affordances results in a sense of individual and/or shared agency for a given action. Where a similar outcome arises across actions or settings, this results in an increasingly stable or recurrent sense of agency, leading to patterns of self-construal, which can vary with culture (Markus & Kitayama 1991).

In this way, culture also significantly influences the experience of agency itself. Above we noted the important distinctions between individual and shared/joint agency and the fact that Doris focuses only on individual agency. There has been a flourishing tradition in social psychology concerning self-construal (Markus & Kitayama 1991; 2003; 2010) – the extent to which people in different cultures vary in the way they represent their sense of who they are by reference to social relationships and group memberships. Markus and Kitayama differentiate between two general, culturally sanctioned ways in which people construe who they are. First, independence is concerned with viewing oneself as separate from others, striving to achieve personal desires, being unique and consistent. Second, interdependence is concerned with viewing oneself as connected to others, striving to maintain harmony, to cohere in one's in-group and with the situation. This offers an individual-level counterpart to broad cultural differences in values between individualism and collectivism. Importantly, Markus and Kitayama note that the distinction ramifies for culturally variable normative models of agency: Independence prompts a model that takes the individual to be the key agent, whereas interdependence prompts a model of conjoint agency (Markus & Kitayama 2010). Voyer and Franks (2014) further suggest that the formation of a dominant independent self-construal is the result of repeated actions requiring predominantly individual agency, whilst the formation of a dominant interdependent self-construal is the result of repeated actions requiring predominantly shared/joint agency. These models suggest different normative tendencies towards construing challenges and opportunities as requiring joint agency ("we") of various kinds, or individual agency. Again, different cultures not only offer different values to enter dialogues for agency, but also qualitative differences in the ways those dialogues function, and their expected outcomes, with important consequences for agency.

To conclude, Doris offers a persuasive yet partial picture of the origin of agency and its relation to self. We applaud his view of the role of social factors and dialogical relations in generating agency. However, we suggest it underplays the importance of individual and shared/joint experience of agency and the complexity and extent of the impact of the social and cultural environments in which these experiences take place.

Learning to talk to ourselves: Development, ignorance, and agency

doi:10.1017/S0140525X17000693, e45

Stuart I. Hammond

School of Psychology, University of Ottawa, Ottawa, Ontario, Canada K1N 6N5.
shammond@uottawa.ca **http://www.stuarthammond.com**

Abstract: Although anti-reflectivism seems to preclude a role for reflection, this dichotomy could be synthesized in a Piagetian developmental framework. Development integrates a role for error and ignorance in reflection, and supports Doris's espousal of valuation, collaboration, and pluralism, and the importance of extrinsic factors to the self.

Doris's (2015b) book provides the groundwork for a constructive solution to an important challenge to traditional moral philosophy. The central problem, which Doris raises here and in his earlier work (e.g., Doris 2002), is that we are generally ignorant of the conditions that influence our actions. We assume that our reflection and reason makes decisions, but that's not how it works. To support this claim, Doris draws on social psychology research (such as the Obedience to Authority experiment, e.g., Milgram 1974). Studies like Milgram's demonstrate that not only do people act in ways that we do not expect (e.g., a majority will choose to shock a protesting research confederate to the point of unconsciousness), but also that small and seemingly morally irrelevant variations in situations (e.g., how far away from the experimental subject the formerly mentioned confederate is sitting) are related to morally relevant variations in outcomes.

Understanding how small situational differences lead to big consequences would seem to require some kind of theory of human nature, but the classic social psychology research does not provide one (e.g., Sarason et al. 1975). Turning elsewhere in psychology, one way to understand how small changes produce big results is to explain contemporaneous human actions through reference to the past, whether ancestral (see, e.g., Tooby & Cosmides 1990) or developmental (e.g., Carpendale et al. 2013). Maybe we get scared of snakes because this helped us survive in the hunter-gatherer era, or perhaps static illusions fool us because of the way we have developed our perceptual

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

abilities in conditions of motion. That said, I know of no evolutionary or developmental account of Milgram, nor am I going to provide one here.

Instead, I'm going to focus on a particular brand of developmental theory in the remainder of my commentary, and mainly focus on Doris's (2015b) definition of reflectivism as the idea that "the exercise of human agency requires accurate reflection" (p. x). I am going to argue that drawing on developmental theory allows us to integrate aspects of reflectivism and anti-reflectivism, in allowing an important role for both accuracy and error in human action. When we contrast reflectivism and anti-reflectivism, we could see these as utterly opposed. But another way is to see these as two sides of the same coin, or as properties emerging out of the same system. Shadow and light emerge from illumination. Sickness and health are the properties of the state of living systems. Although anti-reflectivism can logically be opposed to reflectivism, nevertheless in a developmental psychology approach, where reflection emerges out of action, these could be integrated. This somewhat Hegelian way of doing things hints at the organismic worldviews that motivate many developmental theories (Overton 2006).

Doris (2015b) leaves developmental psychology largely untapped, perhaps because he wants more than mere "causal" agency, i.e., the type of agency that explains organisms moving around in the world (p. 40). However, there are developmental theories that attempt to connect causal agency to more robust conceptions of moral agency (see Sokol et al. 2015 for a review). The most elaborate of these accounts, with all its warts, faults, and difficulties, is the work of Jean Piaget (e.g., Chapman 1988; Piaget 1932/1965; 1963/1936; 1974/1976). A central feature of Piaget's theory is that intelligence is organized action. From a Piagetian perspective, both the six-month-old baby struggling to bring a heavy toy to her mouth and the teenager trying to understand what happened to Michael Brown in Ferguson, Missouri, are exercising intelligence. Where teens differ is in their capacity for reflection. Yet despite the presence of reflection and many other qualitative and developmental differences between the teen and baby, there are common processes in their intelligence, of disequilibrium of their earlier assimilation of the world, and attempts to accommodate to their new circumstances (Chapman 1988).

Piaget (1974/1976) provides a nice example of the complexity of reflection in a study where he asked children and adults to crawl, and then to reflect on and model the process of crawling (i.e., reconstruct the process of crawling). Needless to say, no one had a problem crawling. But when asked to represent crawling, things got messy. Some did this accurately, reflecting that crawling occurs with an "X" pattern, front limb, contralateral back limb, and so on. Others modeled crawling with a "Z" pattern (front limb, other front limb, then back limbs) or an "N" pattern (e.g., left front, left back, then right front etc.). It got messier. When those who had modeled crawling incorrectly were asked to crawl again, some subsequently got it right (i.e., crawled with an X pattern, and then accurately reflected that crawling uses an X pattern). Others, especially young children, demonstrated a kind of parallelism (i.e., continued to crawl one way and model crawling another). And still others *reorganized their crawling to fit their previously incorrect model of crawling* (e.g., began crawling in an N pattern).

What does this example show? To me, it's not clear that even the X pattern is such a great model of crawling. It gets the limb order right, but says nothing about muscles, back posture, and so on. A type of self-ignorant action remains a primary component of how we get about in the world, a conception with which I think Doris could concur. A lot of the time, our reflection captures only part of what's going on. Piaget emphasized that initial reflection most often captured the periphery, such as limb order, rather than the central processes (Hammond 2014). The relation between even these partial models and our actions is complex. Sometimes, as Doris emphasizes, these just seem to run parallel to each other. But, other times we might eventually get it right (er). But then, even an incorrect reflection can influence our actions.

I think this last point is the most interesting one. In some cases our reflections can shape our actions, *even when these reflections are inaccurate representations of the state of the world*. And my subsequently reorganized actions may create a world that more closely resembles what was in error. To some extent, having a reflective conception of myself as a moral person might actually make me one, as I talk politely, and open doors for others. But that conception doesn't do so well when I encounter a Milgram-type scenario. With my understanding in disequilibrium, and a lot of Milgram's (1974) subjects were in such a state, what happens next? There is an interesting dynamic of reflecting, acting, getting things wrong, detecting our mistakes, overlooking them, ignoring them, trying to change, and so on.

In fact, just this very type of question has motivated a great deal of research in moral developmental psychology. Just as Milgram (1974) sought to understand Nazi Germany, so too did figures like Lawrence Kohlberg, who questioned why some people obeyed, and others resisted, and why people change over their life spans (see e.g., Rutland & Killen 2015; Turiel 2008). I will admit, however, that developmental theories such as Kohlberg's have often fallen back on the very type unproblematic reflectivism that Doris aptly criticizes (see Carpendale 2000). The overall point is that this brand of developmental account can problematize reflection as a difficult process (e.g., Campbell & Bickhard 1986) even if they sometime fail to do so.

On the other hand, if reflection were just some kind epiphenomena, which is a position held in at least some conceptions of moral psychology (e.g., Haidt & Joseph 2004), why bother? Why evolve and develop some system of reflection if you have some other system, whether a lower-order emotional processing system, or an evolved modular system, that gets you around in the world? Piaget's system posits reflection as a form of perspective taking, of integrating the dynamic and changing world and people around you (Bickhard 2016). Piaget was adamant that reflection was not the proverbial light that illuminates the dark places of the unconscious, or in his terms, action schemes (Piaget 1974/1976). Reflection is a construction, one that it is invariably incomplete (which is to say always partially ignorant, and wrong), but can also be partially right. That last bit means that although the teenager is unlikely to understand the complex set of factors that lead to what is happening in Ferguson (see e.g., Jones et al. 2015; Tate 2016), her reflection, through her own experiences, talking with friends, through social media, might be very different, and in many ways more accurate, than my mine or Doris's. And perhaps through conversation and perspective taking, all of us might come away with a somewhat more accurate perspective.

My main argument here has been centered on the process of reflection, although I've hinted at how a developmental account support valuation, collaboration, and pluralism. One of Piaget's big sociological contributions was to argue that moral development arises not because of authority or obedience, as Durkheim would have it, but because of mutual respect and collaboration (Carpendale 2009; Piaget 1932/1965; Vaish & Tomasello 2014). To the extent that the Piagetian tradition is developmental, it also incorporates a type of pluralism. We don't judge the toddler in the same way as we judge the teenager.

I'll close by drawing the same issue that Doris raises at the end of his book, that of North America's Indigenous peoples and the personal impact of extrinsic conditions. Doris discusses the United States' Indigenous history, largely in the past tense, remarking how the destruction of extrinsic aspects of Indigenous culture destroyed Indigenous identity, even if people persist. If Canada's history with Indigenous people has been slightly less bloody, it is almost equally awful, relying on residential school systems and an Indian Act to "de-Indianize" Canada's Native population, resulting in a "cultural genocide" (Truth and

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Reconciliation Commission of Canada 2015) alongside death through brutality and disease.

Back in the present, suicide rates among Indigenous youth in Canada are extremely high, frequently described as an epidemic. Research by developmental psychologists Michael Chandler, Chris Lalonde and colleagues explore the connection between extrinsic factors and questions of identity and agency (e.g., Chandler et al. 2003). Their research shows that as Indigenous communities regain governance of institutions such as fire-fighting, health, language, and policing, which are called markers of cultural continuity, suicide rates tend to go down, in some cases to much lower levels than in the European Canadian population. These markers of cultural continuity are extrinsic factors but also support a process of building back a sense of narrative, identity, and agency as youth begin to reflect on their future and their past.

Talking to others: The importance of responsibility attributions by observers

doi:10.1017/S0140525X1700070X, e46

Stefanie Hechler and Thomas Kessler

Department of Social Psychology, Friedrich-Schiller-University, 07743 Jena, Germany.

thomas.kessler@uni-jena.de stefanie.hechler@uni-jena.de
http://www.sozialpsychologie.uni-jena.de

Abstract: This commentary extends Doris's approach of agency by highlighting the importance of responsibility attributions by observers. We argue that (a) social groups determine which standards are relevant and which actors are responsible, (b) consensus about these attributions may correct individual defeaters, and (c) the attribution of moral responsibility reveals agency of observers and may foster the actors' agency.

In his book, Doris (2015b) develops an account of morally responsible agency based on repeated expressions of an actor's value. The book is ambitious and develops a complex argument with a lot of thought provoking points and observations. The author argues that if people exercise agency, they are also morally responsible. He further argues that cognition is socially embedded: People seem to think most efficiently in contexts where other people are around (or one imagines this, or in any case one does not feel isolated). We generally agree with this argument. In the end, Doris endorses pluralism in agency that allows various criteria to account independently for moral responsibility. Any alternative account may simply add to his picture. However, we think that the socially mediated development of (moral) agency should receive much more attention, as we think it is more than a simple addition. The (repeated) attributions of moral responsibility by observers may be much more important than the question of whether people are actually morally responsible agents. We argue that the collaborative nature of the development of agency should be extended to a full social endeavor.

To elaborate this suggestion, we will refer to factors that determine blame and praise and thereby responsibility attributions. These are illustrated by several psychological findings. We will then connect the attribution of responsibility to the problem of actual agency by suggesting that collective (dis-) approval expresses the observers' values and additionally enables actors to learn the adherence of standards.

In order to attribute responsibility to an actor, it may be enough to distinguish roughly between coerced and uncoerced behavior (Strawson 1962). Observers evaluate uncoerced behavior with reference to norms, standards, ideals, or values (further referred to as standards) that are perceived as appropriate in a particular context. These are usually standards that are shared because of a common group membership (Turner et al. 1987). Such groups could be families, moral communities, work teams, or occupational or national groups among others. Observers, usually fellow group members, blame actors for failing to live up to these standards, or praise and admire them for approximating them (Kessler & Cohrs 2008). Hence, they tend to punish deviation from standards (Darley et al. 2000) and admire behavior that approximates high standards (Onu et al. 2015). Most of the time these are group-based standards, which can be found on different levels of inclusiveness. "Morality" may refer to the standards appraised within the most inclusive category, "humans." However, reactions to deviants also allow for group-specific (e.g., cultural) norms, or go beyond humanity and, for example, include animals as proposed by animal rights organizations. Often, observers like to share their evaluations with others, which leads to effective coordination and closeness to the people with similar evaluations (Peters & Kashima 2007).

Like the actor's agency, the observer's attribution of responsibility may be driven by (arbitrary) causal psychological processes ("defeaters"). However, the attribution of responsibility (other than responsible agency as described in Doris's book) is usually performed by several observers. People validate their perceptions and beliefs with reference to their fellow group members (see, for example, Cialdini & Goldstein 2004). Numerous observers that communicate in one or the other way produce consensus, such as conforming to what most people say, exchanging arguments, and correcting oneself (if one notices that one is affected by a defeater). Through this process, some accidental defeaters tend to cancel each other out because people are susceptible to the same defeaters to different degrees, and each observer may even be affected by different defeaters. Thus, mutual social influence suggests that expressions of values by different observers can clarify the appropriate standards more precisely than individuals could do because, as a collective, they are less threatened by individual defeaters.

Various findings show the importance of adherence to group standards: Observers react to particular behaviors with reference to whether they are performed by ingroup or outgroup members, and whether they affect ingroup or outgroup members (Hechler 2016). Group members remember ingroup deviants better than norm-conforming ingroup members and all outgroup members (Hechler et al. 2016). Group members also derogate deviant ingroup members more harshly than deviant outgroup members (Marques et al. 1998). Deviating new group members are treated more educationally, whereas deviating experienced ingroup members receive harsh punishment (Pinto et al. 2010). The nature of the violated standards is also crucial for the assignment of punishment. Deviations from minimal standards (i.e., either-or-standards) lead to harsher punishment (including even social exclusion) than deviations from maximal standards (i.e., gradual standards; Berthold et al. 2012; Fritsche et al. 2009; Kessler et al. 2010; see also sacred values, e.g., Baron & Spranca 1997).

Observers also praise and admire behavior revealing high competence or moral excellence, which causes them to emulate and learn from these admired persons (Onu et al. 2016a) or groups (Onu et al. 2015; 2016b). Such reactions to others' behavior (i.e., punishment and reward) both influence the targets' behavior; the targets tend to adhere more to the standards (Balliet et al. 2011).

The blame and praise by observers attributes (or at least assumes) agency. However, how is it important for actual agency or the development of agency? We think that the assignment of blame and praise is itself valuational because any person evaluating another's behavior expresses her values. With several observers, individual defeaters may cancel each other out (see above) with the effect that their shared evaluation may approximate their actual value. Thus, the assignment of praise and blame is the group's expression of shared values. Deviations from standards tend to invalidate these standards. The communication of blame and praise allows observers to regain consensus

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

and mutual endorsement of the standards. In the face of deviations from standards, people have to express condemnation in order to make the violated standard salient to all group members (Feinberg 1965; Durkheim 1893). In addition, the expression of blame for norm violations demonstrates that group members care about the norms and the group members protected by those norms. Finally, observers blame norm violators to distance themselves from the deed and avoid being associated with such misdeeds. Thus, in some sense the observers show agency when they blame and praise others' behavior because it expresses their values (usually, shared values). They may even express their values without caring too much for actual responsibility of the actors (i.e., they may not go further than differentiating between coerced and uncoerced behavior).

Moreover, we argue that the assignment of blame or praise for misdeeds also affects the actors' agency. Public condemnation indicates, claims, or even fosters group members' exercise of agency. As observers attribute responsibility to the actors, the actors may also perceive themselves as having agency (or an illusion of agency?). For example, children's agency develops by the guidance of sanctions. Agency may be considered an *ability* (that one could learn) instead of a *habit*. Habits denote what people are accustomed to do, whereas abilities include a normative component that denotes what would count as a correct or incorrect thing to do (Millikan 2000). This normative component specifies when we sometimes succeed in expressing our values and when we fail to express them. As mentioned above, praise and blame direct us thereby in the standard's (valued) direction. In contrast, habits could go in any direction, as they are not necessarily corrected by values. Moreover, by such development of ability over time (i.e., agency-training), we become more reliable in expressing our values in particular situations and apply them to more diverse situations.

As an additional mechanism, we suggest that reminders of our responsibility, such as blaming and praising of certain behaviors, activate the concept of personal agency. Activated concepts also tend to produce concept-related behavior (e.g., the belief that one excels in math enhances math performance, Miller et al. 1975). Activated concepts also change cognitive processing characteristics that lead to the enactment of these concepts (Sassenberg et al. 2017). Accordingly, actors who are held responsible may activate their concept of "being responsible." Thus, before acting, they may think twice, activate their main values, and take precautions to make sure that their behavior conforms to their values. Such a reflection of personal values in turn may lead to a stronger connection of these standards in their cognitive system; they may identify with them and thereby behave more in accordance to them. This is also a social process: it not only involves solitary thinking but also social negotiation and training in justifying behavior in the face of others. This may reward careful action, so that people may arrange their environment in order to avoid known "defeaters" (e.g., temptations). Moreover, being held responsible indicates being watched. This enhances objective self-awareness and thereby a person's own standards become more salient.

The social shaping of agency and responsibility may not always work out completely. Some people may be hard to train or unwilling to develop stable "virtues" (i.e., habits to act according to their own and commonly shared standards). However, this may be irrelevant, as others will still hold them responsible (even if this cannot apply literally) and punish them (e.g., go for incapacitation as a last resort). In addition, people may not want to wait until repeated misdeeds manifest the "negative" values of the actor. There may be an asymmetry in that many positive deeds are necessary to manifest positive values of people, whereas one negative deed can be enough to reveal the negative value of an actor. The extremity of the deed may itself be a clear indicator for moral responsibility (Pauer-Studer & Velleman 2011). In such cases, where the social shaping of individual agency or responsibility may be impossible or come too late, the actor can only be made incapable. However, the general practice of collaboratively shaping agency may not be threatened by this because these examples remain exceptions.

In short, the emergence of agency and responsibility is a social process. Talking to others (including blaming and praising) is a particularly efficient way to develop one's own agency and help others become responsible actors.

Grounding responsibility in something (more) solid

doi:10.1017/S0140525X17000711, e47

William Hirstein and Katrina Sifferd
Department of Philosophy, Elmhurst College, Elmhurst, IL 60126.
williamh@elmhurst.edu sifferdk@elmhurst.edu

Abstract: The cases that Doris chronicles of confabulation are similar to perceptual illusions in that, while they show the interstices of our perceptual or cognitive system, they fail to establish that our everyday perception or cognition is not for the most part correct. Doris's account in general lacks the resources to make synchronic assessments of responsibility, partially because it fails to make use of knowledge now available to us about what is happening in the brains of agents.

Our commentary on Doris's significant book focuses on three areas: (1) Doris's claim that cases of self-ignorance, such as confabulation, are common enough to negate our own judgments of why we did things; (2) Doris's inability to give a good account of synchronic assessments of responsibility; and (3) the disconnect between Doris's account and scientific accounts of human thought and behavior.

Self-ignorance. Doris says that human beings are "afflicted with a remarkable degree of self-ignorance" (précis abstract). But while we certainly at times show self-ignorance, there is no absolute metric that allows us to assess the exact degree of our ignorance compared to our self-knowledge. This opens the possibility for researchers, who feed on a steady diet of examples of ignorance, to overestimate its degree. We need to leave open, for example, the possibility that we are dealing not with phenomena that afflict everyone, but with phenomena that only afflict a minority of people, or even a certain personality type. The scope of Doris's skepticism is also broader than it might appear. One sign that we might be overestimating the amounts of ignorance and error is that we have not been moved to enact major changes in folk-psychology to remove dependence on our capacity for self-knowledge. Doris's view seems to commit us not only to being "routinely mistaken" (précis abstract), but also not ever noticing that we are, and attempting to correct it. Doris seems to be neglecting all those times we *aren't* buffoons.

A comparison with the case of visual perception is illuminating. Even though cognitive scientists have cataloged perhaps hundreds of visual illusions that reveal the seams and flaws of our visual system, the vast majority of our visual perceptions during the day are veridical and serve us quite well. Consider our abilities to visually identify one another. Certainly there are many ways in which the brain systems that achieve this miracle can fail, leading to odd syndromes like prosopagnosia. In the everyday sphere, we have all experienced cases in which we visually misidentified someone. But taken against the overwhelming percentage of correct identifications we make so effortlessly and frequently, these misperceptions are rare. This high rate of effectiveness is due to good equipment.

We think serious cases of ignorance or mistaken self-knowledge are somewhat rare because they typically involve errors at two levels. First, a mistaken impression is created. For instance, it occurs to me that I don't really have to pay back that loan from my friend because he seems to be wealthy, when I would just

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

prefer to keep the money. Then, this error is not corrected (this correction could occur because I note my obligation to repay, or I revise my sense of my friend's situation, or I just realize I am being selfish). The first type of error, where I form a mistaken impression of my own motives, is fairly common; the second, where I fail to correct, or at least *where I fail to correct because I cannot correct*, less so. And in cases where we have the capacity to correct for our mistaken perceptions, using our brain's prefrontal executive processes, it would seem we are responsible for them (Hirstein et al. 2018). For example, a color-blind person can correct for his problem by memorizing the location of the traffic lights. Doris's view amounts to saying that the entire upper level that has been designed into our brains, including the executive processes and consciousness itself, is of little use or import. This level functions precisely to correct basic errors of perception or memory, as can be seen in the case of confabulation (Hirstein 2005). This second level tends to only activate when the stakes are appropriately high, so that examples of our failures where they aren't perceived to be high, such as the case of people failing to put money into the office coffee fund, are not showing our cognitive system at its best.

Synchronic assessments of responsibility. Doris argues that moral responsibility for an act depends upon whether the act in question was an exercise of agency (Doris 2015b, p. 159). Exercises of agency, according to Doris, are expressions of the actor's values; attributions of responsibility turn on whether an actor's values are expressed in an act (p. 159). However, this sort of view faces clear epistemological difficulties, as Doris notes: It will frequently be difficult to determine whether someone holds a value, and actions often seem related to multiple values, some of which may be unknown even to the actor. Plus, "values are expressed over time, and can, oftentimes, only be identified over time," and thus "extended observations" may be required to identify patterns to determine if any particular action is of the sort for which an agent can be held responsible. "If one focuses on isolated events, diagnoses may falter" (Doris 2015b, p. 162). In the end, attribution of responsibility may require first that "a pattern of cognition, rationalization, and behavior emerges, and that pattern is best explained as involving the expression of some value"; and second, a determination that a particular action expresses that value (p. 164). But why in a revolutionary era of neuroscience assume that we must remain forever locked outside the mind and brain of the subject? Doris's account involves the cognitive sciences, but only those that focus on behavior and outward from there, to society. We suggest that connecting his knowledge of the psychological research with neuroscience, via cognitive neuropsychology, would greatly help resolve the epistemic problems involved in discerning what exactly someone's values are.

As it stands, Doris's theory indicates that synchronic assessments of responsibility are often impossible. However, the most common and important responsibility attributions are synchronic. Take, for example, criminal verdicts. Judges and juries do not, and ought not in most cases, focus on past behavior as a means to indicate responsibility for a particular crime.[1] A criminal court is asked to determine whether a defendant held a particular mental state and whether this mental state is causally related to the criminal harm. Such canonical cases of responsibility attribution are considered so secure we use them to deny defendants' liberty and even life. If Doris's theory is correct, and responsibility assessments rest on extended investigations into a person's values, then it would seem our current system of generating verdicts and punishing offenders is likely to attribute responsibility to persons when it has not been proven they deserve blame.

Doris indicates that he is a pluralist about responsibility, and thus "sympathetic" to the possibility that there may sometimes be warranted attributions of other types of responsible agency, including reflectivist agency (Doris 2015b, p. 174). However, he also feels that a pluralistic account must place dialogic agency in an "appropriately prominent" position (p. 175). To vindicate the thrust of his theory with regard to criminal verdicts, Doris should provide an account of how a synchronic act must be related to dialogic agency. Further, this account must explain how a synchronic act can be seen as an expression of such agency without an exhaustive review of the agent's history. But if this were possible, then it would seem that Doris's requirement of "extended observations" would, in most cases, be unnecessary because a less burdensome, synchronic assessment would suffice.

In a similar vein, the reactive attitudes, which Doris acknowledges are important first indicators of responsible agency (Doris 2015b, pp. 23–24), are typically generated in synchronic cases without information about character. They depend on the brain's mindreading (or theory of mind) capacities, through which we attribute motives behind a person's actions, sometimes using fairly few behavioral cues. If these motives are selfish, for example, a strong negative reactive attitude will follow. In the criminal law, we feel stronger condemnation where an agent directly desired criminal harm (committed the act "purposely" under the U.S. Model Penal Code) than in cases where an agent merely ought to have known there was a risk of substantial harm (committed the act "recklessly").

It isn't clear that Doris's weakly proposed pluralism, which encompasses his dialogic view and reflectivism (Doris 2015b, p. 174), can generate many of the synchronic responsibility assessments made in the criminal law. As Doris argues, many culpable actions do not seem connected in the right way to reflective judgments, which are often confabulated. Thus, if extensive investigation of dialogic agency is not done, on what grounds are criminal verdicts generated? For example, in a case where the fire was due to the building owner's forgetting to check the functioning of the water sprinklers, a synchronic assessment of the defendant's conscious mental states with regard to the criminal harm will not secure a responsibility assessment. In our view, only an account that provides a synchronic assessment of capacity for responsible agency, where that capacity is more expansive than just the capacity for conscious reflection, can ground criminal verdicts of negligence.

Personal versus subpersonal. As we noted, Doris chooses to keep his analysis at the personal, rather than the subpersonal level, by using information largely from social psychology. But sometimes, simple knowledge of the person's brain can clear things up. For example, Doris notes that "the valuational account says if your action properly expresses your values, it's an exercise of agency, regardless of whence your values came" (Doris 2015b, p. 30). But what about someone with Tourette's whose outbursts do express his values, but not in a way he wanted? Or a person whose sleepwalking actions do express his values, but are horrible, and which he would never do when awake? In both of these cases, responsibility does not seem to rest with the actor, due to volitional incapacity, despite the alignment of the action with the actor's values. If we could "see" the actors lack of control via evidence of brain function (or dysfunction), we might correct mistaken assessments of responsibility.

Doris searches everywhere for help in attributing psychological states such as motives, including other people (the dialogic part of this theory), except in neuroscience. There is useful information at the subpersonal level, from neuroscience, cognitive neuropsychology, and from historical neurology, that is vital to gaining a full understanding of the relevant phenomena. Neuroscience can provide valuable data regarding synchronic assessments of responsibility. For instance, it might be able to tell whether an action is "done habitually" (i.e., what the neuroscientists call an action done "in routine mode") or done as a result of conscious reflection, which involves quite different and more extensive brain processes. While Doris avows materialism, it is difficult to see how his theory, as stated, can be put into stark, materialistic terms. What concrete things, states, processes, and events do claims about "values," "desires," "plans," "self-awareness," and "the exercise of agency" refer to? We are not done with the project of building a theory of responsibility until we can do that.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

NOTE

[1] Federal Rule of Evidence 404(b)(1) states that "Evidence of a crime, wrong, or other act is not admissible to prove a person's character in order to show that on a particular occasion the person acted in accordance with the character."

Getting by with a little help from our friends

doi:10.1017/S0140525X17000723, e48

Enoch Lambert and Daniel C. Dennett
Center for Cognitive Studies, Tufts University, Medford, MA 02155.
Enoch.lambert@gmail.com Daniel.dennett@tufts.edu
http://ase.tufts.edu/cogstud/dennett/

Abstract: We offer two kinds of constructive criticism in the spirit of support for Doris's socially scaffolded pluralism regarding agency. First: The skeptical force of potential "goofy influences" is not as straightforward as Doris argues. Second: Doris's positive theory must address more goofy influences due to social processes that appear to fall under his criteria for agency-promoting practices. Finally, we highlight "arms race" phenomena in Doris's social dynamics that invite closer attention in further development of his theory.

Doris conducts a master class for psychologists on how to extract value from the philosophical debates, and for philosophers on how to use empirical work in psychology to inform their theorizing. In both endeavors, one has to learn how to take the declarations with more than a few grains of salt, which Doris applies judiciously. We heartily endorse what we take to be a major lesson: What we learn from science, while sometimes shocking, need not destroy our confidence in our own practical agency. Rather, by informing our understanding of our agential strengths and weaknesses, science can guide us in discovering and strengthening those practices that foster our agential powers. Of special note is his case that self-ignorance can be crucial to our projects of building and expressing our central values, showing how accurate reflection can actually undermine agency in some situations. He has also done the study of practical reason a great service by setting up a framework for exploring its socially scaffolded nature. In our comment, we aim to contribute to that ongoing project. While we believe Doris is right about the largely social nature of agency, we raise some questions about the skeptical force of the psychology he cites against the role of accurate self-knowledge in our deliberations. We also urge that his own "collaborative-negotiative-dialogical" framework faces significant threats from social psychology – more so than acknowledged.

Doris's critique. First, we wish to question the strength of the case Doris mounts for *global* skepticism regarding the role of accurate self-knowledge in our deliberations. We are more concerned about the size of experimental effects and their implications for everyday decision making than Doris is. It is instructive to recall the reason why so much psychology focuses on surprising effects. Vast swaths of common wisdom concerning self-knowledge prevent psychologists from so much as attempting to confirm things like whether people tend to be accurate about whether they prefer $1,000 to a pin prick, or social praise to ridicule. Finding a new way of generating small, surprising effects may be rewarded in psychology, but it is not clear whether or how the common lore of everyday psychology that psychologists never bother to investigate is undermined by it.

Doris (2015b) dismisses the importance of statistically small sizes partly by saying that known "goofy influences" on behavior indicate an ocean of unknown ones; and partly by saying that such influences may "aggregate" in ways that medical interventions can (p. 64). Our own speculative mechanics of goofy influences suggest a different lesson. If "eyespots and pronouns are in the mix" (to use Doris's nice phrasing), then humans are likely assailed by goofy influences *continuously* (p. 64). The priming and automaticity literatures from across psychology suggest no principles for ruling out much of anything as potentially goofy influence. But if this is so, how do we manage to hold it all together? Why are we not driven every which way by the onslaught of disparate priming stimuli? And how are we able to come by the amount of common knowledge of human psychology that we do? Why can we predict so well what others will do based on "typical" perceptions and desires (which we also attribute to ourselves)? When predicting what the drivers of other cars on the road will do, we justifiably pay no attention to which images on which billboards they recently saw, or the content of the radio advertisement they are hearing, or whether their vehicle interior is leather, or. . . . It isn't that we are in a position to rule out such things ever having some influence on how they drive, whether at a micro-level, or, on occasion, at a life-altering level. But our attributions are sufficiently reliable enough of the time so that it makes no sense to let such influences trigger general skepticism of our usual interpretive and predictive capacities. Similar considerations apply to our own case. It would be silly, for instance, to decide to live as close as possible to the market simply because it would minimize the amount of goofy influence encountered every time we need to do our shopping.

Moreover, Doris ignores the prospect of a gradient between goofy and not-so-goofy, to go along with his valuable gradient between explicit self-reflection and the sort of automatic self-monitoring that gets us relatively gracefully through the day. The fact that pictures of watchful eyes should nudge more honest coffee transactions is striking, but not so striking or upsetting as the non-fact – we wager – that pictures of bicycles or rooftops have the same effect. Doris's richly detailed account of actual decision making suggests that in the real-time hasty triage involved in all but the most portentous moral decisions, a "subliminal" hint about being observed and caught could be just enough to bias the choices made without the choosers' noticing.

Next consider one of the roughly third of test-subjects who detected the switches in the moral choice blindness experiment Doris cites (2015b, p. 139; see Hall et al. 2012). What should such a subject conclude upon learning the results of the experiment? That she got lucky? Why would that be more reasonable than to conclude that, for whatever reason, she was more attentive (or cared more, or . . .)? *Perhaps* she should conclude that her capacity to recognize her own moral positions is more susceptible to error than she would have thought, and so she should keep watch. But it doesn't seem reasonable to conclude that she should be an outright skeptic of her ability to recognize her own morality. And, in general, we urge that individual variation in susceptibility to goofy influences not be swept aside as so much noise. Why is it that goofy influences do not affect some subjects in any given experiment? Are some people who are less susceptible in specific experiments more generally resistant to goofy influences? If so, why? Can any pattern at all be detected in failure to succumb to goofy influence? It seems that such possibilities remain live empirical hypotheses to be ruled out (or in!) rather than assumed. Until we know more about the mechanics of goofy influences, it seems rash to let them *completely* undermine the role of accurate reflection in our deliberative decision making.

Doris's positive framework. Given Doris's conservatism about our everyday attributions of agency and responsibility, it is surprising that he uses psychotherapy as a model for how collaboration and dialogue can facilitate agency. In the history of agential responsibility, psychotherapy has been around for a blink of an eye, and has been employed by a sliver of agents. So it is at best a device for highlighting what aspects of our common practices actually do facilitate agency. Dialogue and "positive alliance" are the agency-facilitating aspects of beneficial psychotherapy highlighted by Doris. But both phenomena are also present in collaborative enterprises where anti-agential forces often prevail. We review some below, but we encourage Doris to say more about what lessons to take from psychotherapy, as well as

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

suggest additional cultural models for understanding what facilitates agency.

Doris argues that "agency requires . . . mutual influence" (2015b, p. 148), the kind of influence involved in negotiating our relationships via exchange of "rationalizations" (in Doris's non-pejorative sense) (p. 153). While he makes a nice case that, at least some of the time, our own self-ignorance plays a necessary supporting role in such influence, we are less sanguine. For such influence is subject to forces which often do not facilitate the expression of values. Some means of social influence are often simply irrelevant or arbitrary from the perspective of our values. (Robert Cialdini's 2008 classic *Influence* reviews the evidence for many, as well as the fact that they have been exploited far longer than they have been documented by social psychology.) Other forms that may *sometimes* be congruent with our values, such as reciprocity (Cialdini et al. 1975) or deference to authority (Milgram 1974), are also easily exploited to influence in ways that are not at all reflective of our values.

Several types of group processes can drive us toward actions that are in conflict with our values. Processes involving groupthink (Baron 2005; Janis 1982), group polarization (Moscovici & Zavalloni 1969; Sunstein 2009), intergroup phenomena (Sherif et al. 1961; Tajfel et al. 1971), power dynamics (Keltner et al. 2008; Galinsky et al. 2006), etc., can all influence people to behave in ways they wouldn't endorse in non-group or other-group settings. In all such cases, negotiated rationalizations are a part of, or are influenced by, these very processes. Indeed, "*Animal Farm* phenomena," whereby groups or even whole societies become what they usually, or used to, condemn, come about partly through processes that, as far as we can tell, have not yet been excluded from Doris's category of negotiated rationalizations. To be sure, Doris (2015b) insists that the "right" kinds of social processes are needed for agency (p. 125). But it is not yet clear how to specify those processes or how often they occur. Take the case of Milgram's obedience experiments. Doris (2015b) points out how they demonstrate that the presence of some dissent enables more (p. 119). Comrades *can* help us act out our values. But note that the primary experimental paradigm itself falls, so far as we can tell, under the rubric of a collaborative social negotiation, complete with rationalizations (there is a sense in which the experimenter and subjects formed what Doris calls a "positive alliance" – the situation was set up as one in which they team together for the sake of advancing science). How can we know which social groups and processes undermine our agency? Should classic social psychological results be taught in schools (sounds good to us)? What about *reflection* on the results of our myriad interactions?

Doris's excellent demonstration of the complexity and variation found in everyday (responsible) decision making nicely exposes the problem with the traditional philosophical *isms* that Doris so patiently analyzes – and abandons: They all tend to be static and absolute, laying down presumably eternal policies of self-control that may look good on paper but nobody could live with. The fact is that human decision making has always been something of an arms race (in the evolutionary sense), with novel techniques being introduced, identified and warned against, refined, further disarmed, etc., and all memorialized in the world's folktales and literature, from the country mouse and city mouse to Othello and "hidden persuaders." The arms race is intensified and accelerated by human reflection itself, because insightful critics and other observers expose the various ploys and vulnerabilities, and because we all have the desire to persuade others of our own values, we take on the lifelong goal of honing our talents in the game of reason-giving, trying to hold our own as responsible agents who can protect themselves from goofy (and other baleful) influences. Among the social scaffolds invented by arms races are measures like strict liability laws, which have the effect of emphasizing "due diligence" in areas of particular risk of harm by removing in advance otherwise reasonable excuses. Laws criminalizing the technically benign (concealed weapons, drug paraphernalia) or even "intent" (from conspiracy to hate) are also part of the ongoing arms races that reflect the power of human reflection on social scaffolding.

Sometimes we become entangled in meta-strategizing and second-guessing that can stultify us, turning us into accomplished boxers whose footwork and feints dazzle our opponents while we never land a punch. As Doris explains, too much attention to how we are doing can prevent us from being at our best. Most of the time, the task is made easier by benign social scaffolding: we treat each other with "the benefit of the doubt" and this trust is itself a wise policy, provided that the balance is favorable. But there is no foolproof way of preventing us from falling into bad company, and then no pure "individualist" policy can be endorsed. As Doris puts it, "In point of fact, very little of what a person does, be it good or bad, is entirely up to the person herself" (2015b, p. 170).

Agency is realized by subpersonal mechanisms too

doi:10.1017/S0140525X17000735, e49

Neil Levy

Department of Philosophy, Macquarie University, Sydney, NSW, Australia.
Uehiro Centre for Practical Ethics, University of Oxford, Oxford OX1 1PT, UK.
neil.levy@philosophy.ox.ac.uk

Abstract: John Doris argues that, when behaviors are caused by processes that we would not endorse, our agency is defeated. I argue that this test for defeaters is inappropriate. What matters is not what we would but what we should endorse. The subpersonal mechanisms he identifies as defeaters enable us to track and respond to reasons. They realize agency, rather than defeating it.

There is, as John Doris (2015b) emphasizes, extensive evidence that human agents often act in ways that are influenced by mechanisms that respond to features of the world without the agent being aware of the relevant mechanisms or of how they respond to the features. Often, too, these mechanisms work in ways that the agents would not endorse on reflection. Doris suggests that this evidence poses a major problem for the justified ascription of moral responsibility to agents like us. Our vaunted capacity to reflect, deliberate, and make decisions is threatened by the existence of what he calls "defeaters." *Talking to Our Selves* is a systematic response to this challenge. The first part of the book identifies the problem; the second offers a solution to it. In this review I will focus on the first half. While there is much to recommend the account of agency Doris develops in the second, I will suggest that the problem it aims to solve is largely illusory.

A defeater is a cause of a decision or action that would not be recognized as responding to reasons in its favor by the person, were she aware of its influence (Doris 2015b, pp. 64–65). Consider the ballot order effect, for example. There is extensive evidence, cited by Doris, that being at the top of the ballot gives a candidate an electoral advantage. But people wouldn't recognize *being at the top of the ballot* as a genuine reason to favor one candidate over another. When their choice of candidate is influenced by ballot order, it is influenced by mechanisms the workings of which the person would not endorse if she became aware of them. When "the causes of her cognition or behavior would not be recognized by the actor as reasons for that cognition or behavior, were she aware of these causes at the time of performance, these causes are *defeaters*" (pp. 64–65). It is a defeater of what Doris calls agency (which he understands as a capacity of agents to act such that their actions reflect their values), and therefore of moral responsibility.

Many of the processes causally involved in our decisions and actions satisfy Doris's definition of defeater; were we aware of them, we would not endorse them. But the standard provided by Doris's test is not the appropriate one to identify genuine defeaters of agency and responsibility. It does not matter, for these purposes, what we *would* endorse. What matters is what we *should* endorse. Many of the processes that Doris's tests identify as defeaters of agency are better understood as helping to realize agency, whether or not we would endorse them.

On the most plausible account, or family of accounts, morally responsible agency consists essentially in the capacity to recognize and respond to reasons, including moral reasons (Fischer & Ravizza 1998; McKenna 2017). In fact, Doris himself seems to assume such an account, at least as a necessary condition of moral responsibility, which is why he worries that defeaters are such a serious problem for us. They defeat agency by bypassing or overwhelming our capacity to recognize and respond to reasons (2015b, p. 52). But there is no reason to identify the capacity to recognize and respond to reasons with a set of processes that we would endorse were we to become aware of them. What matters is whether the processes actually enable us to track and respond to reasons, not whether we would endorse them were we aware of them. And very many, perhaps the overwhelming majority, of the processes that Doris identifies as defeaters are better seen as realizers of our agency than as defeaters of it.

Many of the processes that are supposed to be defeaters of agency are evolved dispositions. We have these dispositions (to prefer the first presented candidate, for instance) because they were adaptive in our ancestral environment. And in the main they were adaptive because they enabled us to track and respond to reasons better than we would have done were we to rely on slow, effortful, resource intensive, domain-general reasoning. While we live in vastly different environments today, a large proportion of these processes continue to track reasons. They do so whether or not we would endorse them.

Consider, for example, our disposition to prefer the default option when choosing the settings for everything ranging from insurance policies through to organ donation (Johnson & Goldstein 2003). Perhaps we would not endorse this disposition on reflection. But the disposition is likely adaptive, for a range of reasons. First, the default option may be the default for a reason; that is, it may be because it is the best option (or at minimum a satisfactory option) for most people that it is the default. Its selection as the default may reflect its endorsement by those who designed the policy (who are often in a better position to pick the best option than the consumer). Of course, defaults may be chosen arbitrarily or for bad reasons, but if they remain the default over time, the agent can often be confident that it is not a bad choice (by the standards prevailing in her group). If a very large number of people have faced the same decision before her, and the default has remained the default over time, then it is unlikely to be contrary to mainstream prevailing values. Here is one point at which the socially embedded conception of agency Doris defends in the second half of the book indeed provides part of the solution to the problem he addresses: the processes that he thinks of as defeaters are often designed to rely on features of the social environment.

Whether designed by nature or acquired in development, the suite of subpersonal mechanisms that cause our behavior are typically adaptive, and typically they are adaptive because they enable us to track and to respond to reasons. Very often, regardless of whether we would endorse them, they do a better job of tracking reasons than the kinds of processes we would endorse on reflection (like slow, effortful, conscious deliberation). They may do a better job because they are fast and frugal, allowing for good enough decision making in conditions in which speed is at a premium or the expected marginal benefit of engaging in effortful deliberation is too small to justify the expenditure of time and resources needed for effortful deliberation. Often, however, they do a better job than conscious deliberation would do, were the agent to take the time to engage in it. When the data are noisy, for instance, we often do better to employ a simple heuristic (of the kind embodied in subpersonal mechanisms) than to employ conscious deliberation, because tracking a few cues yields better results (Gigerenzer 2008). Even under conditions conducive to deliberation (when the problem is tractable computationally, deliberation is capable of outperforming subpersonal mechanisms and the decision is important enough to justify the investment), conscious deliberation may often lower decision quality relative to the employment of heuristics (Wilson & Schooler 1991). People may endorse conscious deliberation and reject subpersonal processes, but the second may do just as good, or better, a job at tracking reasons and thereby enabling agency.

Even when subpersonal mechanisms cause us to make choices in ways that fail to track reasons, they typically do not bypass our agency. Consider the ballot order effect again. While sometimes candidate order conveys information about the quality of the candidate (Marcinkiewicz 2014), often it does not. In many electoral systems, voters face a choice between candidates assigned ballot order by lot or alphabetically. Random allocation does not correlate with candidate quality, and alphabetical order is unlikely to (though there might be very weak indirect effects; perhaps having a name with an initial letter that occurs early in the alphabet leads to more opportunities to speak in school environments, for example). It is noteworthy, though, that ballot order effects make a significant difference to the choices of two groups of voters: low information voters and those who are nearly indifferent between options about which they are knowledgeable. Now, while the ballot order effect can be expected to make a difference to the choices of many people in these two groups, it does not thereby bypass their agency (that is their capacity to express their values in their actions). For those who are indifferent between the option chosen and another, both of which they're informed about, each choice expresses their values just as well as the other, so the effect does not make a difference between expression and its absence. The choices of low information voters may not express their values, but the primary reason for this is not because the choice is influenced by ballot order: It is because they don't know enough for their choice to express their values. The choice is not a worse expression of their values than the one they would make were they to reflect more. Low information voters susceptible to ballot order effects are also indifferent between options because they don't know enough about them to make a choice. When the ballot order effect influences them to make a choice, that choice is not a worse expression of their agency than one they might have chosen had they reflected.

Of course, the subpersonal mechanisms that orient us toward some considerations and away from others sometimes lead to suboptimal behaviors. Such mechanisms may fail to track reasons because a mechanism that was adaptive (by tracking reasons) in the environment of evolutionary adaptiveness may no longer function to track reasons in our very different environment. How often such mismatches between environments cause the bypassing of agency is an open question. The extent of bypassing is limited by two factors. First, social forces often work to ensure that our dispositions to choose do not depart very significantly from satisfactory choice by altering the environments in which we choose (again, a default option that is not satisfactory for most people will likely be culturally selected against). Second, subpersonal mechanisms are not deployed blindly; rather, they are more likely to be deployed when they are appropriate (Todd & Gigerenzer 2007). Subpersonal mechanisms may also simply misfire. Such a mechanism may still appropriately be regarded as partially constitutive of morally responsible agency: Conscious deliberation, too, is prone to misfiring in unpropitious circumstances.

Doris's principal target in *Talking to Our Selves* is the view he calls reflectivism. Reflectivism is the view that cognition and behavior is agential only when it is preceded or accompanied by reflection on how to behave. As Doris argues, reflectivism is hard to square with the general drift of the evidence from

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

cognitive science: If only actions that are appropriately ordered by reflection count as instances as morally responsible agency, then there are precious few instances of such agency. The conflict between reflectivism and the view urged here, according to which agency is pervasively realized by subpersonal mechanisms that are opaque to introspection, may be reduced by the recognition that consciousness is genuinely important for flexible response in novel situations. It is important, not because conscious deliberation is powerful, but because consciousness is the gateway to global availability to the subpersonal mechanisms that realize agency (Levy 2014). Behavior does not need to be ordered by reflection to be agential. Rather, the suite of mechanisms that constitute us also make us genuinely reasons-responsive agents.

ACKNOWLEDGMENTS
Partially supported by a grant from the Wellcome trust (WT104848/Z/14/Z).

Acting without knowledge

doi:10.1017/S0140525X17000747, e50

Heidi Lene Maibom
University of Cincinnati, Cincinnati, OH 45221.
heidi.maibom@uc.edu
http://www.artsci.uc.edu/faculty-staff/listing/by_dept/philosophy.html?eid=maibomhi&thecomp=uceprof

Abstract: I question whether psychological effects that an agent is unaware of can express her values and, if they can, whether this allows us to hold her responsible in the range of cases that we would like to.

Responsibility always was a difficult issue, and it is not getting any easier. When we knew relatively little about mental processes, it was easier to think of the forces that might interfere with our agency in such abstract terms as 'laws of nature' and in relatively external ways (environment, upbringing). As we learn more about psychology, these interfering forces – or 'effects' as they are usually called – seem more personal and perturbing. We are subject to Order Effects, Bystander Effects, the Better-Than-Average Effect, and so on. At any one time, it seems, we have no way of knowing the extent to which our actions are influenced by any of these operating conditions of our psychological machinery. This is apt to lead to considerable skepticism about freedom and responsibility. Doris makes a compelling case for the problem in his *Talking to Our Selves*, but is also kind enough to offer what he takes to be a solution. The account is a slight modification of the Frankfurtian idea that central to agency (or personhood) is our ability to form second-order volitions. Doris calls such volitions 'values.' Values are desires that we have put in the driver's seat, Doris says. We are only responsible for actions that express our values. So far, so good.

Doris embraces a rather strong form of skepticism about what we can know about the causes of our actions. Because of the way our minds work, our actions are sometimes, if not always, influenced by things other than our values. Can we be responsible for such actions? Yes, Doris says, as long as our actions can be seen as *expressing* those values. This opens up the real possibility that actions that express our values are not actually *caused* by our values. Indeed, Doris seems to welcome this conclusion, even if it does not sit comfortably with his insistence that we are only responsible for actions that are *self-directed*: "self-ignorance often functions to effect self-direction, and its absence can be an impediment to agency" (Doris 2015b, p. 144). How is this possible? We are presented with a range of cases where an agent's actions that are *caused* by unconscious mental influences end up furthering his or her well-being or projects. For instance, the Illusion of Control is the illusion that you have more control over events than you actually do. But suppose that this illusion prompts you to work harder at saving a faltering relationship that you value. Now this influence *enhances* your agency, Doris claims, because it is more likely to help you achieve your goal (of saving the relationship). Even if it is the Illusion of Control that drives your effort – unbeknownst to you – your actions still *express* your valuing of the relationship.

But there are a number of problems with this solution. First, it is unclear how your working on the relationship *expresses* your valuing the relationship if, in fact, it is the Illusion of Control that's driving your actions. Typically, what is expressed is part of the cause of the expression. Take, for instance, emotions. When we express an emotion, the emotion or its eliciting conditions is the cause (depending on your view of emotions). The causal chain may be more or less direct. In Doris's case, however, there may be *no* causal chain from your values to the expression of them. But can something that has nothing to do with your valuing the relationship express it? Perhaps the idea is this. Works of art can express things that did not cause them, such as desperation, joy, or anxiety. Fine. But can such forms of expression be linked to responsibility? Recall that *self-direction* is central to responsibility, and this is analyzed in terms of values. But in the cases Doris mentions, the person does not direct anything by way of her values; it is her psychological quirks (the Illusions of Control, say) that cause her to act as she does. This means that her working on her relationship *coincides* or *is consistent* with her valuing the relationship. But this is hardly sufficient for her being held responsible for working on the relationship *if* this requires self-direction.

Hang on, you might say. Was it not part of Frankfurt's point that a willing addict could be held responsible for taking the drug because his second-order volition was one of embracing his addiction, *even if* it was the addiction that caused him to take the drug? Doris's position seems to be no different. The following counterfactual appears to be at work: *had* the Illusion of Control not influenced her actions, the agent would *nonetheless* have acted in the same ways to salvage the relationship. This may seem acceptable. Now we are holding an agent responsible not so much for the actual action she performs, but for an action that *she would have performed* had she been free to do so. This may be as good as it gets for agency. We should note, however, that determining what someone *would have done* is tricky. Philosophers are adept at constructing examples that make compelling cases (e.g., Frankfurt 1969), but reality tends to be messier. I don't think we can really suppose that had our subject *not* been under the control of the Illusion of Control, she would nonetheless have performed the very action she performed because she valued her relationship. I don't see how we could possibly know that. Neither could she.

Even if you find the solution palatable for cases such as our relationship example, how should we think of other instances of actions influenced by psychological effects? Suppose that the effect in action *does not* fit with your values. Take the Bystander Effect. You see a man fall over on the street. Nobody helps him. Neither do you. But this does not express a value of yours. Indeed, you value helping others. Now you seem like the unwilling addict. You are not responsible for your action because you would not have performed it had you not been subject to the Bystander Effect. If this way of modeling Doris's ideas is right, then we ought to analyze all human action as instances of Frankfurt-style addiction. Only if our values happen to coincide with the forces that influence our actions are we responsible for these actions. This is made clear by supposing that you value *not* helping others. Others should be self-sufficient, and you don't have *any* responsibility to come to their aid (so you believe). If you now act under the influence of the Bystander Effect you *are* responsible because this action coincides with your values.

One, no doubt unintended, consequences of this view of things is that people are rarely, if ever, responsible for wrongdoing. Most people do not think of themselves as evil or even averagely bad. They think of themselves as basically decent people. They are unlikely to have put desires in the driving seat that are the sorts of values that we see expressed in wrongdoing. The drunk

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

driver does not value drinking over killing another human being. Is there a way of describing his action in such a way that it makes sense to hold him responsible? Perhaps he values drinking over the safety of others? Presumably, he would not admit to holding such a value. Do his actions nonetheless *express* such a valuing? In an earlier work, Doris maintains that we may be self-deceived about the values we hold (Doris 2002). Our actions may *reveal* that we hold values that we would not openly endorse or that we have not considered. We might, therefore, say that the drunk driver *is* responsible, because his drinking and driving expresses a disregard for the safety of others. It is important to note, though, that this determination can only be made on the basis of a *pattern.* A one-off drunk driving offense is not enough to show that the agent values drinking over the safety of others. This suggests, then, that only repeat offenders can be held responsible for their wrongdoing. The rest of us are quite likely off the hook. That doesn't sit right with me. Even if I do not hold a value of not helping people in need, I can nonetheless be blamed for not helping a person in a Bystander scenario. The point is not that I express my disregard for the person by my inaction, but that I should, and could, have known better (Maibom 2014).

Whether or not we get problems with blaming (i.e., too little of it) on this view of responsibility, it certainly seems that we end up with a pretty radically curtailed number of actions that people can be held responsible for. The problem ultimately speaking is this: either valuing is a substantial process arrived at through a significant amount of cognitive work, in which case many of our actions do *not* reflect our values, *or* most of our actions reflect our values, in which case 'valuing' means little more than, perhaps, giving in to a desire. At different points in the book, Doris appears to lean in one direction, then in another. Suppose I decide to kill the squirrel that has decimated my strawberry patch. I get an air gun and shoot it. Does this action express a value that I hold (low valuing of squirrels, say)? I doubt it. I am an ardent wildlife supporter. I give a lot of money to such causes. I feed wild birds, and cry inside when I see a flattened squirrel by the road side. I certainly enjoy strawberries and growing my own food, but not above everything else. Indeed, were I to sit back and consider the value of a squirrel's life relative to the, say, 40 strawberries that I'm likely to harvest, I would see the squirrel's life as more important. Moreover, the shooting is hardly part of a pattern of disregard for squirrels. And so the conclusion seems to be that I am not responsible for killing the squirrel. But does this really seem reasonable? If I *am* responsible, though, this cannot be because of any real *value* that I hold.

To conclude, whereas I agree with Doris about much of what he says in his book – that psychological effects present a challenge to responsibility and that agency is deeply intertwined with our interactions with others – his solution to the problem does not satisfy me. I cannot see how an action caused by effects that have no *internal* relation to an agent's values could possibly *enhance* her agency. At best, it can sidestep it. But even here we face problems, such as that it seems to lead to our rarely being responsible for our actions, particularly for the bad things we do.

Talking to others' selves: Why a valuational paradigm of agency fails to provide an adequate theoretical framework for moral responsibility, social accountability, and legal liability

doi:10.1017/S0140525X17000759, e51

Tobias A. Mattei

Neurosurgery and Spine Specialists, Eastern Maine Medical Center, Bangor, ME 04401

tobiasmattei@gmail.com
https://www.emmc.org/Providers/Mattei,-Tobias-A-,-MD.aspx
https://www.researchgate.net/profile/Tobias_Mattei

Abstract: In this commentary, I highlight the importance of a proper discussion of the pragmatic implications of John Doris's paradigm for allocation of personal responsibility proposed in his new book *Talking to Our Selves*. By employing some classic concepts of the American common law tradition, I discuss why Doris's valuational understanding of agency fails to provide an adequate framework for moral responsibility, social accountability, and legal liability.

In his new book *Talking to Our Selves*, John Doris (2015b) provides a comprehensive analysis of the age-old discussion about the existence of truly self-determined behavior, agency, and moral responsibility (Cary 2007). Although throughout the book the author mainly focuses on the underlying philosophical assumptions and empirical data from the psychology literature that may justify his position (which he characterizes as anti-reflectivist, valuational, and dialogic), I believe that a proper discussion of the pragmatic implications of the paradigm defended by Doris is of paramount importance.

As once contended by Richard Weaver (1948) in his famous work, *Ideas Have Consequences*. Therefore, if the philosophical position defended by Doris in his new book is to be taken seriously, it is expected that, similarly to past works on the issue (e.g., Duff 1990), the proposed conceptual framework should be able to transcend the purely theoretical realm, ultimately bearing significant practical implications to other social sciences, including the field of legal studies. Here I provide an evaluation of the generalizability (or lack thereof) of Doris's valuational paradigm of agency for justification of moral responsibility by applying it to several distinct areas of the classic American common law tradition.

American tort law has assumed, since its inception, a clearly distinctive approach regarding culpability. The hallmark of such a departure from the strict liability paradigm sponsored by the English common law tradition is the classic *Brown v. Kendall* (1850) case decided by the Massachusetts Supreme Judicial Court. The circumstances of this legal dispute involve a fight between two dogs belonging to different owners, both of whom ended up getting involved in the animal fight in an attempt to separate them. During the effort to do so, the defendant beat the dogs with a stick and, in the process, accidentally wounded the plaintiff in the eye, causing him a severe bodily injury. The plaintiff brought suit against the defendant for assault and battery. During the arguments, the plaintiff's lawyers presented a persuasive argument requesting the Massachusetts Court to employ a strict liability standard for judging the defendant's actions based on a classic case of the English common law tradition decided in 1466, *Hulle v. Orynge*, best known as the "Case of Thorns" (King's Bench 1466). In this case, when attempting to retrieve thorns that dropped onto Plaintiff's property, the defendant entered the plaintiff's private field, ultimately damaging some crops. In its final decision, the English court ruled that, although the defendant had a reasonable justification to enter the plaintiff's property, he was nonetheless liable for trespass. According to the Case of Thorns' precedent, one who voluntarily performs an act that results in damages to another is responsible for the damages even if the act was itself lawful in nature. The American judges involved in the *Brown v. Kendall* case, however, proposed a quite different standard than the one advocated by their English peers centuries before. They basically ruled that if an accidental casualty arises from a lawful act, no tort action can be brought by the Plaintiff unless he is able to demonstrate that the Defendant acted with lack of "ordinary care." Such judgment ended up becoming a legal cornerstone of the American tort law, ultimately generating a clear binding precedent that established the requirement of fault for all tort cases. Such a legal position is embodied in the classic Latin principle: *nulla poena sine culpa* (no punishment without fault).

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Although a superficial analysis might lead some to conclude that Doris's valuational paradigm would be compatible with the U.S. tort law standards simply because his criteria would lead to the same result as the official *Brown v. Kendall* judgment, there is a deep degree of incompatibility between them. The official criteria employed in the *Brown v. Kendall* case for allocating moral responsibility and, therefore, for tort liability, involved two distinct prongs: the lawfulness of the activity under question and the presence (or absence) of negligence as defined by the criteria of the 'ordinary care employed by the reasonable person.' Neither one of these is properly taken into account by Doris's valuationist paradigm. In other words, while according to the U.S. tort law it is possible for individuals to be held legally liable for inflicted damages either because the original activity under question was illegal or because there was a substantial degree of negligence involved in the action which ultimately led to the harm, according to Doris's valuational paradigm neither one of these criteria (i.e., legality or negligence) would be enough to allocate moral responsibility as long as the individual caused the harm unintentionally.

Another example of how Doris's valuational paradigm for allocation of moral responsibility fails to address important legal issues (in this case with regard to criminal law) is the hypothetical case described in chapter 7 (Doris 2015b). In this example a father, after an exhausting work day, forgets his child inside a locked car under a dangerously hot temperature. According to Doris, ascribing moral responsibility to such a parent would require invoking the concept of 'strict liability.' In criminal law, strict liability consists in a standard through which the individual can be found liable for a crime based only in the *actus reus* (the guilty action) regardless of the underlying state of mind (i.e., *mens rea* or guilty mind) (Carson & Felthous 2003). According to the traditional common law standards, strict liability would never be considered as a reasonable standard to be applied to the action described in this hypothetical case. In fact, in the American legal tradition, strict liability (which represents a form of absolute legal responsibility for an injury that can be imposed on the wrongdoer without proof of carelessness or fault) is mainly restricted to specific areas of the tort law, such as product liability (in which there is a collective social interest in protecting consumers against defective products) or in the case of minor criminal offenses or misdemeanors (such as traffic violations), which usually do not carry major social stigmas. The illustrative example presented by Doris would be better described as involving an essential element of negligence. According to the Model Penal Code, which has served as the theoretical basis for criminal statutory laws in several U.S. states, there are four different *mens rea* descriptors of culpability that can be used to qualify an individual's action (Robinson & Grall 1983): "purposely," if the criminal act was the conscious objective of the agent's conduct; "knowingly," if the result was not the agent's primary or conscious objective, yet he could be practically certain that his conduct would secondarily cause that harm; "negligently," if the agent's conduct involves a gross deviation from the standard of care that the reasonable person would observe in the same situation; and "recklessly," if the agent was aware that his conduct would involve a substantial and unjustifiable risk of causing the harm. It is opportune to recall that even when an individual acts purposely (e.g., an intentional crime) the law does not require (or seem to care about) the presence (or absence) of an adequate awareness of the underlying motives for such an action. Therefore, similarly to Doris (although for completely different reasons), most legal scholars would also simply reject the reflectionist position on the issue of moral responsibility and avoid any discussion of the reality of 'incongruent parallel processing' on grounds of irrelevance.

The main point of our discussion here is not to ascertain (or to deny) that the supposed parent who abandoned his child in the dangerously warm car acted negligently. That is the role of the prosecution attorney who will construct his legal argument on the specific details of the case under question. However, it seems clear from this example that by employing Doris's valuational paradigm, such a parent could never be considered morally responsible (and, therefore, legally liable) for his action apart from a strict liability basis, as the parent clearly did not act according to his "deepest and most cherished values." In his book Doris also proposes a simple 'test' to determine if an action has been performed according to the agent's core values and, therefore, if the individual can be held morally accountable for such an action. "I say a behavior is an exercise of morally responsible agency when the actor is self-directed . . . when the nicotine addict guiltily succumbs to craving and lights up his behavior is not self-directed" (précis, sect. 5, para. 3). According to this rule, the parent cannot be held morally responsible (and, therefore, legally liable) for his negligent action, as he was not self-directed in his behavior, which can be ultimately traced to a tragic concurrence of unfortunate external circumstances (the exhausting workday, the dangerous warm weather, etc.) as well as a unintentional personal lapse.

A final cornerstone dogma of the American common law tradition that seems to be incompatible with Doris's valuational paradigm for allocation of moral responsibility is the concept of vicarious liability. The doctrine of vicarious liability represents a foundational doctrine of the Western legal tradition that finds expression in several distinct tenets, such as the concept of command responsibility in military law, the concept of employers' liability (i.e., doctrine of *respondeat superior*), the concept of principals' liability (through which the owner of an automobile can be held vicariously liable for negligence committed by a person to whom the car has been lent), and the concept of parental liability (through which parents can be held liable for tortious actions committed by their children due to their own negligent behavior expressed as a failure to supervise them). For example, according to the above-mentioned doctrine of *respondeat superior* (a Latin expression meaning "Let the master answer"), an employer is liable for the tortious acts of employees performed within the course of their employment (Gared 1983). In the United States, due to the Supreme Court decision in *Pinkerton v. United States* (1946), individuals can even be considered vicariously liable for crimes committed by others if the offenses were performed in furtherance of an unlawful agreement or conspiracy (Alex 2008). The words of U.S. president Ronald Reagan (1987), in a speech in the Oval Office while commenting on the Iran-Contra scandal, expresses the depth and the importance of the *respondeat superior* doctrine: "First, let me say I take full responsibility for my own actions and for those of my administration. As angry as I may be about activities undertaken without my knowledge, I am still accountable for those activities. As disappointed as I may be in some who served me, I am still the one who must answer to the American people for this behavior. And as personally distasteful as I find secret bank accounts and diverted funds – well, as the Navy would say, this happened on my watch." Nevertheless, according to Doris's valuational paradigm for allocation of moral responsibility, it would be impossible for an individual to be held, as in the Reagan's example, vicariously morally responsible for the action of others.

In summary, Doris's valuational paradigm of agency seems irreconcilable with key principles of the American common law tradition, ultimately failing to provide an adequate conceptual basis for allocation of moral responsibility, social accountability, and legal liability. In the same way that skepticism regarding moral responsibility may only be a feasible philosophical option at the personal level (Waller 2011), ultimately failing to offer any reasonable alternative for the development of a viable legal code, if the interesting theoretical scheme developed by Doris fails to provide that type of practical and coherent conceptual framework that the professionals at the law school across the university campus require for proper exercise of their duties, there is a significant risk that such a well-written and delightful book may represent a mere exercise of intellectual digression, a classic epitome of the speculative literary output of isolated scholars "talking to our selves" inside our own ivory towers.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

A limited skeptical threat

doi:10.1017/S0140525X17000760, e52

Joshua May
Philosophy Department, University of Alabama at Birmingham, Birmingham, AL 35294-1260.
joshmay@uab.edu **www.joshdmay.com**

Abstract: Doris argues that our choices are heavily influenced by forces that we wouldn't count as genuine reasons. This unsettling conclusion is motivated by a debunking argument so wide-ranging that it isn't foisted upon us by the sciences. Doris sometimes seems to lower his ambitions when offering instead a skeptical hypothesis argument, but that conflicts with his aims in the book.

John Doris (2015b) argues forcefully and eloquently that human thought and action aren't quite what they seem. He deftly points to empirical research that suggests that our actions are commonly influenced by a wealth of unconscious and, importantly, *unseemly* factors: "Many studies identify causes of behavior that are not plausibly taken as reasons for behavior" (p. 43). One of his favorite examples is the finding that people appear to cheat less when there is a depiction of eyes watching them (e.g., Bateson et al. 2006). Few people would happily say "I did it because of the eye spots" (Doris 2015b, p. 43).

Human actions, Doris concludes, are often driven by unconscious and unreflective processes that amount to "defeaters." These are influences "the actor is unaware of, and would not recognize as a reason justifying the behavior, were she so aware" (Doris 2015b, p. 52). Doris doesn't quite give this view a label. He just associates it with a "skeptical threat," which he thinks we can avoid by adopting his own preferred theory of agency.

Defeaters are supposed to be particularly damaging to a "reflectivist" tradition, which holds that human thought and action are normally guided by "accurate reflection" on one's own mental states. Now, various other philosophers and scientists, including myself, likewise believe that the human mind isn't so reflective (e.g., Watson 1975; Arpaly 2003; Seligman et al. 2016; May, forthcoming). However, Doris also doubts the "accurate" bit, suggesting that we often aren't motivated by what we'd regard as genuine reasons. Here I want to suggest that the skeptical threat has been overstated, and yet a weakened version isn't enough for Doris's ambitious purposes.

A debunker's dilemma. So far the skeptical threat looks to be motivated by a genealogical debunking argument, in which some beliefs or other attitudes are allegedly influenced by illicit processes. Nietzsche and Freud, for example, famously attacked ordinary moral and religious beliefs as being influenced by wishful thinking, egoism, and rationalization. More recently, some philosophers argue that ordinary moral beliefs are unjustified because they have been too heavily shaped by extraneous evolutionary forces (e.g., Joyce 2006). Whatever the targeted attitudes, such debunking arguments intend to reveal that the attitudes are problematic because they're *in fact* substantially influenced by unseemly forces (see Nichols 2014).

Genealogical debunking arguments can be made to work, particularly when informed by the relevant empirical research, provided they aren't too wide-ranging (Kumar & May, under review). Empirical evidence can reveal that some of our decisions are influenced by arbitrary factors, but it's more difficult to establish that most of our behavior is so influenced. Indeed, there is plenty of evidence that many of our choices are influenced by good reasons (see e.g., discussion in Batson 2011; Miller 2013; Seligman et al. 2016).

With evidence on both sides, it looks like we ultimately have to do the hard work of determining what does drive most of our decision-making. Yet Doris can't just show that our choices are slightly influenced by arbitrary factors. That would leave room for reflective direction or for being influenced by unreflective yet appropriate factors. We must ask: Are (a) *most* of our choices (b) *substantially* driven by (c) genuinely *arbitrary* factors?

My own view is that a close examination of the empirical literature suggests otherwise (May, forthcoming). Certainly, some of the influences on our choices are truly *unwelcome and arbitrary*. For example, our choices certainly shouldn't be determined by racial or gender bias, irrelevant feelings of disgust, or the mere order in which information is presented. But a comprehensive look at the literature, including meta-analyses, suggests such arbitrary influences are often rather small (see e.g., Oswald et al. 2013; Landy & Goodwin 2015; Demaree-Cotton 2016). This leaves plenty of room for being motivated primarily by the right reasons. For example, feeling queasy from food poisoning *might* make one think stealing is *slightly* worse than one would judge otherwise (May 2014). But the reason most people don't embezzle from their employer is because they think it's unfair, harmful, disrespectful, or just plain immoral.

Other choices are *substantially* influenced by various factors, such as group size, ambient smells, being in a hurry, similarity to a victim, and honor codes (see e.g., Latané & Nida 1981; Carlson et al. 1988; Batson 2011; Ariely 2012). But these forces, while often powerful, aren't necessarily something we'd reject as non-reasons once we examine the effects in more detail. Consider, for example, being in a hurry, being in a good or bad mood, or being reminded of one's moral commitments. We may be happy to cite these as genuine reasons for either helping or not helping a stranger in minor need. Imagine: "Why didn't I stop to help that man pick up his dropped papers? I was in a hurry and I'm just not in the mood to talk to anyone right now." Even when a stranger's situation appears to be dire, one has good reason not to help if one infers that no real help is needed because someone else will do it or everyone else who hears what's going on isn't helping. Even if one recognizes another is in serious need, feeling compassion may be a good reason for help. I may empathize more because the victim and I share a similar background and gender, but that may just draw my attention to a relevant reason to help (e.g., he's in serious need).

This isn't just ad hoc whack-a-mole. There may well be a general *dilemma* here for wide-ranging debunkers like Doris: Influences on many choices tend to be either substantial or arbitrary but not commonly *both*. Indeed, some influences may turn out to be neither substantial nor arbitrary. Eye spots, for example, aren't necessarily arbitrary, for they may serve as a moral reminder that draws one's attention to reasons for being fair and honest (same goes for honor codes and the like). And one meta-analysis of 25 studies suggests the eyes-effect is small and quickly diminishes (Sparks & Barclay 2013). Either way, at least one of the conditions for a debunking argument isn't met.

Skeptical hypotheses. Doris might avoid this dilemma by reframing the ambitions of his argument and thus its explanatory burdens. At one point he does explicitly model his approach on skeptical hypotheses in epistemology meant to undermine knowledge of the external world (Doris 2015b, p. 65). Such perceptual skeptics argue that your evidence would be the same if you were hallucinating or being fed fake experiences by a Cartesian evil demon. Because you can't rule out the possibility that your experiences are systematically deceiving, you don't know there is a physical world beyond your senses (Brueckner 1994). The idea is not at all that this grand skeptical scenario is *actual*, only that it's *possible*.

Doris accordingly thinks he has only to raise the mere possibility of a skeptical scenario. There is a "large, odorous, and ill-tempered animal under the awning of agency," he writes, and thus "for all one knows, any decision may be infested by any number of rationally and ethically arbitrary influences" (2015b, p. 64). For Doris, the "critical question concerns not how often defeaters should be thought to obtain, but how their presence can be ruled out" (2015b, p. 68).

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

However, if this is the form of argument, then we didn't need all of the empirical evidence. Imagination alone can generate hypothetical scenarios in which it systematically seems we're motivated by good reasons though we're not. Moreover, while philosophers have long been fascinated with this form of argument, it's not necessarily because they find it compelling.

Perhaps Doris's idea is that his skeptical hypothesis argument should be more persuasive because there is some positive scientific evidence that the skeptical scenario is actual (compare Sinnott-Armstrong 2006). Nevertheless, skeptical hypothesis arguments make no claims about the actual genealogy of the relevant mental states. The only empirical claim in such arguments is about the actual character of one's evidence – namely, that it can't rule out the skeptical scenario – but this isn't a claim about the source of one's attitudes (May 2013). So, given Doris's explicit and extensive appeal to evidence that our choices and decisions are in fact influenced by arbitrary factors, he seems to be offering a wide-ranging debunking argument.

Doris presumably requires a debunking argument anyway for his purposes. Skeptical hypothesis arguments lead to a sweeping denial of knowledge. Perceptual skeptics conclude that we don't know there's an external world; Doris concludes we don't know our behavior is defeater-free. Such negative conclusions cut both ways: We neither know that we are, nor that we aren't, perceiving an external world or acting for good reasons. Doris, however, aims to establish the positive claim that much of our behavior is in fact influenced by defeaters. His rejection of reflectivism, for example, relies on knowledge of the influences on our actions. As he says, the argument "gets its bite from a family of empirical observations indicating that reflection does not [in fact] play the sort of role in self-direction that reflectivism supposes" (2015b, p. 33). Moreover, Doris's own dialogic theory of agency is motivated by such claims about the actual springs of human action. So it seems Doris needs to do more than raise the specter of malodorous influences on our choices.

Conclusion. In the end, I think Doris is quite right that much of our behavior is determined by unreflective processes – certainly more than commonsense suggests. But only some of our choices are substantially determined by unsavory causes. Of course, we should still pay close attention to these. Even small biases can add up, generating large social problems. What's much less clear is whether small or rare biases warrant overhauls in our conception of human agency and moral responsibility.

A related proposal: An interactionist perspective on reason

doi:10.1017/S0140525X17000772, e53

Hugo Mercier

Institut des Sciences Cognitives – Marc Jeannerod, CNRS UMR 5307, 69675 Bron, France.

hugo.mercier@gmail.com

https://sites.google.com/site/hugomercier/

Abstract: This comment introduces the interactionist perspective on reason that Dan Sperber and I developed. In this perspective, reason is a specific cognitive mechanism that evolved so that humans can exchange justifications and arguments with each other. The interactionist perspective significantly aligns with Doris's views in rejecting reflectivism and individualism. Indeed, I suggest that it offers different, and maybe stronger arguments to reject these views.

Doris's admirable book brings to bear on the question of moral agency insights from many disciplines. In particular, Doris relies on various psychological findings to question standard, individualistic theories of moral agency, and offers an alternative in which moral agency is partly the result of social interactions. In this comment I would like to draw attention to a recently developed theory of reason (Mercier & Sperber, 2017) and suggest that it not only significantly converges with Doris's conclusions, but also pushes them further in several directions.

Dan Sperber and I have developed a theory of human reason, offering a new understanding of what reason is and of its evolutionary functions. In this theory, reason is a cognitive mechanism – a module, or set of modules – that is dedicated to the evaluation and production of reasons. Although it is tempting to equate reason so understood with the System 2 of dual process theories, there are in fact significant differences.

We suggest reason is "just" another inferential mechanism, one that does not supersede all of the other mechanisms, and one that shares most of their properties (by contrast with the quasi-homunculus which System 2 tends to turn into). As other inferential mechanisms, finding and evaluating reasons is, in most cases, quasi-effortless and automatic (think of how hard it would be to avoid understanding a reason as a reason when confronted with one, or to avoid thinking of reasons when your views are challenged), and largely intuitive. This last point is especially important: Reason delivers intuitions about the quality of reasons. When we look for reasons, or encounter reasons offered by others, we typically have an immediate intuition regarding their quality (i.e., how much support they offer whatever representation they are offered as a support of). Like other intuitions, these intuitions are opaque to us. In some cases we can provide reasons for our intuitions about reasons, but the chain must stop pretty quickly. Developing further reasons to support even the most seemingly mundane reasons has provided philosophers with job opportunities for centuries. For instance, most people might provide as a reason for believing that Everest is the tallest mountain on Earth that they have read that in an authoritative source. But why is that a good reason? Because this authoritative source is usually right . . . but why is that a good reason? And so forth. To put it differently, even the most explicit reasons rely on implicit premises that cannot all be made explicit.

In this framework, reason is just one cognitive mechanism among a great many others, and it is these other mechanisms that are responsible for the vast majority of our actions and beliefs. It would be impossible for reason to offer an accurate account of the functioning of these mechanisms – after all, we are talking about the human mind, the most complex computational mechanisms ever evolved (that we know of). Instead, reason can at best focus on some of the factors that might justify our actions or beliefs. It could not plausibly give an exhaustive account of these factors.

This suggests that Doris's criticism of reflectivism is too generous. Reflectivism, in a strong form at least, is psychologically implausible for two reasons. The first is that our minds are simply too complex for one very small part of them (reason) to be able to understand the whole of the rest. The second is that reasons are always partly implicit, so that reflectivism is necessarily partial (we might be able to provide some reasons for our actions, but not to justify why these are good reason, etc.).

With this in mind, the cases of incongruent parallel processing brought up by Doris as arguments against reflectivism seem less striking. If we accept that the production of reasons as justifications is necessarily deeply imperfect, then the reasons provided by people in cases of incongruent parallel processing are not much worse than many reasons that might seem, a priori, fine. Consider two voters. One voted for Candidate Creepy even though his name was second on the ballot, but has only a weak preference for this candidate. The second voter also chose Candidate Creepy, but would have voted for Candidate Normal had she been first on the ballot. Both give the same reasons for justifying their choice. It might be that the set of psychological factors that led to these two choices are quasi-identical. Our first voter might simply have had a very slightly stronger

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

preference for Candidate Creepy. In neither case are the positions on the ballot the most relevant factors. Even for the second voter, the relevant factors would be those that led her to have no strong preference between the two candidates (otherwise she would have picked whatever candidate was strongly favored, irrespective of the order on the ballot). As a result, when both voters present the same reasons to support their choice, the first is being barely more accurate than the second. Singling out cases of incongruent parallel processing might be persuasive because the examples are striking, but it skirts the largest problem looming over reflectivism.

The theory Dan Sperber and I developed thus bolsters Doris's attack on reflectivism. It also considerably strengthens his case regarding the importance of interaction for establishing moral agency. Our theory suggests that human reason would have evolved chiefly to serve two related functions, which are both social. The first (and most relevant here) would be to justify our actions, and evaluate others' justifications, so that we can evaluate one another more accurately. The second would be to offer arguments for our beliefs, and to evaluate others' arguments, so that we can communicate more efficiently.

The social functions of reasons might help explain why reflectivism is an intuitively appealing theory of moral agency. The justifications people offer suffer from the flaws described above: They are partial, and they necessarily contain implicit premises. But that does not stop them from being helpful justifications in a social setting. Following on Doris's example, imagine that our voters offered as justifications for voting for Candidate Creepy that they think he can restore America's lost greatness. This is socially helpful in several ways. First, as Doris reminds us, the voters suggest that they are rational agents, who should thus be sensitive to reason. Second, it provides some insight into the actual factors that caused their actions. Even for the voter who was influenced by the name order, the reason provided might have been one of the actual factors that led him to support Candidate Creepy. Third, it binds the voters to some extent, so that they should pay a social cost if they started deriding another candidate for wanting to restore America's lost greatness.

If reasons can play this role, it is because people intuitively understand that "restoring America's lost greatness" can be a reason for supporting the candidate whose platform that is. They might disagree, but they understand – everybody has the trivial intuition that making something we like good again is a good thing. That's why people do not realize that part of the reason is implicit: They intuitively fill in the blanks. In turn, this is what makes reflectivism intuitively plausible.

If our account might help explain why reflectivism is intuitively plausible, it also resolutely makes of reason a social mechanism. Reasons are for social consumption. One reason our account might not converge fully with Doris's collaborativism is that it considers also cases in which the relation between the parties is more adversarial. Although we do not expect the exchange of reasons to be productive when the parties have no common incentives whatsoever (as in a poker game, say), reasons are most helpful when full collaboration cannot be expected either. It is when people interact with others they do not fully trust that they can most benefit from the exchange of reasons. If agents trusted each other fully, they should be charitable in understanding each other, and would have less need to justify their actions. By contrast, in most cases our default when interpreting others' behavior is to be uncharitable (Malle 2006), and justifications can help revise initially uncharitable estimates. This is one of the reasons why we dubbed our account of reasoning "interactionist" rather than collaborativist (even if we agree that the exercise of reason has to be mostly cooperative to be evolutionarily plausible).

As I hope this brief comment makes clear, Doris's account and ours converge in the rejection of reflectivism and individualism, and can likely strengthen each other.

Why value values?

doi:10.1017/S0140525X17000784, e54

Samuel Murray

Philosophy Department, University of Notre Dame, Notre Dame, IN 46556.
smurray8@nd.edu
https://philosophy.nd.edu/people/graduate-students/sam-murray/

Abstract: Doris argues that an agent is responsible for her behavior only if that behavior expresses (a relevant subset of) the agent's values. This view has problems explaining responsibility for mistakes or episodes of forgetfulness. These problems highlight a conceptual problem with Doris's theory of responsible agency and give us reasons to prefer an alternative (non-valuational) theory of responsible agency.

Between 2010 and 2014, U.S. fire departments responded to nearly 166,100 home fires *per year* that involved some sort of cooking equipment. The National Fire Protection Association reports that 49% of these fires were the result of "unattended cooking," where people forgot that they turned on the stove (Ahrens 2016, pp. 39–41). While these house fires are costly, they aren't nearly as tragic as Forgotten Child Syndrome. On average in the United States, since 1998, 37 children die *per year* of heatstroke because their parents forgot them in hot cars (Weingarten 2009). In 2016, 39 children died of heatstroke in the United States because of parental forgetfulness (Null 2016).

If you think these cases are exceptional, consider the last time that you forgot to call a friend on her birthday, pick up something from the store, or attend a meeting. Unless you're a living saint, chances are you forgot one of these things recently.

I mention these cases because they raise problems for Doris's valuational theory of morally responsible agency. Here, I want to state precisely the phenomenon that these cases capture, the problems that they raise for valuational theories of responsible agency, and alternative (non-valuational) theories that avoid these problems.

First, let me summarize briefly Doris's valuational theory of responsible agency. Doris claims that in order for you to be responsible for some action or outcome, then you must have exercised agency in so acting or in bringing about that outcome. Further, one exercises agency when one's behavior is caused (non-deviantly) by one's values or a suitable subset of one's values (Doris 2015b, pp. 24–26). Thus, Doris finds a tight connection between moral responsibility, agency, and value expression. This is a valuational theory of responsible agency, so-called because the theory locates responsible agency in valuative psychological states that figure causally in action production (cf. Bratman 2000; Frankfurt 1988; Watson 1975). The contrast to a valuational theory is a reasons-responsive, or capacitarian, theory of responsible agency (see Fischer & Ravizza 1998; Nelkin 2008; Vargas 2013; Wolf 1990). Capacitarian theories state that responsible agency is a function of the possession and exercise of agential capacities. I'll return to capacitarian theories after noting some problems with Doris's valuational theory of responsible agency.

Consider, again, the cases mentioned at the outset. Leaving the stove on, forgetting the child in the back seat, and missing the meeting are all examples of mistakes or *slips* (cf. Amaya 2013; 2015; 2016). Briefly, an agent slips when she acts intentionally in a way that is contrary to her preferences. Slips can result from a number of different failures. For example, a slip might result from a failure to form an appropriate intention (as in the case of forgetting to call your friend on her birthday). Or a slip might result from failing to implement an intention at the right time (as in the case of leaving children in hot cars) or failing to maintain an intention for the appropriate duration (as in the case of leaving the stove on too long). The important point is that slips result from mistakes or failures that are also intentional actions *and* that conflict with an agent's overall balance of preferences (cf. Amaya 2013, p. 569).

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Sometimes when an agent slips, there might be value expression. I wouldn't turn on the stove unless I loved hot, home-cooked meals. I might also be responsible for the slip (or for the resulting fire damage to the kitchen). But in the case of responsibility-apt slips, the value expression does not explain the responsibility. I'm not responsible for the kitchen fire because I love home-cooked meals. This point generalizes to other responsibility-apt slips: The valuational theory of agency consistently misidentifies the target of the responsibility attribution in cases of slips. When some agent slips (and is responsible for the slip), the agent is responsible for an omission (e.g., forgetting the birthday, forgetting the stove, failing to check the back seat) and that omission does not express a relevant subset of the agent's values.

This is problematic for Doris because he claims that there is a tight connection between responsibility and value expression via exercises of agency: "Responsibility is typically associated with agency" (2015b, p. 155) and "archetypal exercises of agency are expressions of the actor's values" (p. 159). Something has to go here. Either we can have exercises of agency without value expression, or we can have responsibility without exercises of agency.

Doris anticipates this objection somewhat, citing the existence of "candidates for responsible behavior that aren't exercises of agency," which candidates are the slips mentioned above (2015b, p. 154). Doris responds with the claim that we shouldn't hold people responsible for slips. And if you don't like that response, he thinks that a weaker claim is available, namely that "responsibility is *typically associated* with agency" (2015b, p. 155; emphasis mine). Thus, even if we concede that people are responsible for some of their slips, we can still maintain that in most cases an agent will be responsible for some bit of behavior only if some subset of the agent's values causes the behavior in the right sort of way.

There are two problems with this. First, we regularly hold people responsible for their slips (if you're in a significant relationship with someone, then you can see this firsthand: Just forget about an anniversary or special day and see how that goes for you). So we should prefer theories of responsible agency that preserve this part of our practices to theories that don't (this aligns with Doris's practical conservatism on p. 158). Second, it's not clear to me that responsibility-apt slips are rare. After all, how often do people forget things or make mistakes? That's partly an empirical question, but the answer is not obvious enough to warrant asserting that responsibility is *typically associated* with agency (at least if we understand exercises of agency as expressions of values). This shows that Doris cannot dismiss slips as easily as he supposes.

With respect to slips, capacitarian theories have a clear explanatory advantage over their valuational counterparts. This is because when agents slip (or when agents slip in morally significant ways), there is generally some responsibility-relevant capacity that the agent fails to exercise in some scenario where she could and should have exercised that capacity. For instance, I have recently argued that in some cases of morally significant slips, the agent exhibits a failure of vigilance that makes it appropriate to blame her for those slips (Murray 2017). Insofar as vigilance is a responsibility-relevant capacity, vigilance can figure into a capacitarian theory of responsible agency that captures our intuitive reactions to cases of morally significant slips.

Doris, however, concedes that agency and responsibility may come apart at times. This leads him to endorse pluralism with respect to agency, where diverse psychological processes can support exercises of agency. Thus, in those instances (like slips) where agency and responsibility come apart, Doris thinks that non-valuative agential structures can undergird exercises of agency (2015b, pp. 174–75). Retreating to pluralism, however, does not adequately address the issue. The challenge from slips highlights a *conceptual* problem with valuational theories of morally responsible agency, not just an extensional inadequacy. The conceptual problem is that while responsibility, agency, and value expression typically coincide, they fail to exhibit the kind of explanatory relations that the valuational theory predicts.

If that diagnosis is correct, then we should ask what reason we have to accept a valuational theory. Doris offers two reasons for thinking that exercises of agency just are (or are constituted by) value expressions. The first is that taking exercises of agency to be expressions of values explains why we hold some entities responsible and not others (2015b, pp. 24–25). For instance, we hold human beings responsible but not cats, tornadoes, or carpets. Human beings have values and express them in action, whereas the other three do not. But the capacitarian can explain this distinction. Human beings have certain capacities that underwrite certain expectations about how we will behave. Animals, natural disasters, and inanimate objects lack these capacities, so we do not hold them morally responsible for their behavior (Murray 2017, pp. 508–509). The second reason that Doris offers is that creatures that lack values altogether seem strange targets for responsibility ascriptions (2015b, p. 25). The capacitarian can say two things here. First, the lack of values might signal a lack of responsibility-relevant capacities constitutive of moral agency (e.g., such an agent might lack prospective memory or cognitive control). Second, the capacitarian might concede that values structure responsibility-relevant capacities while still maintaining the primacy of capacities in settling the appropriateness of certain responsibility ascriptions.

So neither of Doris's reasons provides rational pressure to accept valuational theories over capacitarian theories, in part because the capacitarian can explain both of the phenomena that Doris cites without making any untoward adjustments to her theory. What, for Doris, tips the balance in favor of the valuational theory?

The real value in the valuational theory is that it answers the skeptical challenge that Doris poses. Roughly, Doris (2015b) argues that empirical evidence suggests that people are not sufficiently reasons-responsive in the way that capacitarians suppose (cf. pp. 51–52, 171). The empirical evidence does not cast similar doubt on the role that valuing plays in our lives. In particular, the phenomenon of choice blindness provides evidence that people are not responsive to reasons. If people were responsive to reasons, then we would expect to find reversed statements in choice blindness experiments detected at higher rates (p. 139). And given the fact that choice blindness can occur even in morally significant contexts (Hall et al. 2012), the capacitarian theorist is at a significant disadvantage. *This*, I think, is the real reason why Doris plugs for a valuational theory of responsible agency.

While choice blindness studies provide some evidence against capacitarian theories of responsible agency, the evidence is not decisive – especially given the replication crisis in psychology (cf. Doris 2015b, pp. 44–49; I mention the crisis as someone that accepts Doris's methodological commitment to empirically informed theorizing, so I don't mean to be dismissive of scientific data). Also, the survey questions that Hall et al. (2012) use to study choice blindness are somewhat complicated. Would the effect remain if the questions were simple moral statements like 'Torture is wrong'? Surely people's judgments with respect to those statements would be stable. So why don't people notice switching in Hall et al.'s experiments? It could be that people don't really understand the statements. Or it might be that people just aren't paying enough attention to what they're saying and doing. Surely *something* must account for the difference between seemingly stable judgments with respect to simple statements like 'Torture is wrong' and the emergence of choice blindness with respect to complex statements such as this example from Hall et al. (2012, p. 2): "Large scale governmental surveillance of e-mail and Internet traffic ought to be forbidden as a means to combat international crime and terrorism." Here's another, more speculative, interpretation. Perhaps with complex moral statements about the permissibility of government surveillance, people are ambivalent. They can think of reasons

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

that support mutually exclusive positions. The initial judgment accords with the reasons that the subject finds salient at that moment. When the experimenters reverse the answer and ask for justification, perhaps the switched answer makes salient the considerations in favor of the competing position. If that interpretation is correct, then it explains why complex statements but (supposedly) not simple statements generate choice blindness.

In any case, the capacitarian can interpret the choice blindness studies within the confines of her theory. And if she can furnish an adequate explanation of choice blindness, then we're left without a reason to prefer a valuational theory of responsible agency to a capacitarian theory. And given the conceptual inadequacy of valuational theories (highlighted by cases of slips), I think there is significant pressure to reject valuational theories for capacitarian ones.

ACKNOWLEDGMENTS
Thanks to Laura Wells and Maria Altepeter for discussion of various parts of this commentary. Special thanks to Santiago Amaya for written comments on an earlier draft.

Acknowledging and managing deep constraints on moral agency and the self

doi:10.1017/S0140525X17001182, e55

Laura Niemi[a] and Jesse Graham[b]
[a]Department of Psychology, Harvard University, Cambridge, MA 02138;
[b]Department of Management, David Eccles School of Business, University of Utah, Salt Lake City, UT 84112.
lauraniemi@fas.harvard.edu jesse.graham@eccles.utah.edu

Abstract: Doris proposes that the exercise of morally responsible agency unfolds as a collaborative dialogue among selves expressing their values while being subject to ever-present constraints. We assess the fit of Doris's account with recent data from psychology and neuroscience related to how people make judgments about moral agency (responsibility, blame), and how they understand the self after traumatic events.

In *Talking to Our Selves,* Doris (2015b) grapples with the problem of whether and how people ought to be considered morally responsible agents when they do not seem to be able to access accurate accounts of the reasons for their own behaviors. He spends a good portion of the book gathering findings from psychology experiments to demonstrate that people are better at fooling themselves than knowing themselves. We act with a number of arbitrary and ridiculous influences pulling the strings, and when we attempt to explain ourselves by looking inside, we wear rose-colored, and awfully smudged, glasses. Thus, taking psychological science seriously, Doris positions himself as skeptical about people's ability to exercise morally responsible agency. However, he contends that people may sometimes exercise morally responsible agency to the extent that their behaviors express their values. How could we possibly know when a person's behavior expresses his or her values? Given rampant self-deception and self-ignorance, we're asked to be wary of what probably strikes as a good signal: a person's willingness to mobilize a verbal defense of his or her behaviors. The *dialogic* or *collaborativist* aspect of Doris's theory addresses worries about how to precisely determine when a person has acted according to his or her values, by pointing out that the continual cognitive penetration of people's evaluative judgments by external forces – including, importantly, the value-driven questioning of others that occurs in dialogue – renders their values not truly their *own*. The collaborativist view of agency hinges on a collaborativist notion of the *self* in which what individuals count as valuable for the self depends on what other people count as valuable.

Moral responsibility, in turn, is determined through exchange and negotiation of reasons, in an unfolding, collaborative conversation. Ostensibly, as happens in negotiations, for a matter to be considered settled on both ends, both parties will trade off pleading and conceding until they can peacefully move on from it. So it is more than okay to consider people self-directed, value-driven agents when they, for example, initially claim ignorance about moral permissibility or when they are unable to articulate their position, in addition to easier cases, such as when they appear to be squeamish about making value judgments or taking a stand. By equating agency with negotiation, a collaborativist view of moral agency "trades in uncertainty" (sect. 2, para. 12), and is normatively neutral. Interestingly, one way Doris's account of the exercise of moral agency maintains neutrality is that it accommodates an interpretation that is congenial with people's interests in social justice (moral agency is participatory action), but also maintains throughout a deeper, sometimes unsettling, message about constraints (moral agency is inevitably never truly up to one person).

As psychologists pursuing the scientific study of the unruly domain of morality, we consider Doris's empirically based philosophy of moral agency an endlessly thought-provoking accomplishment. In the spirit of collaborative conversation, we offer some more data from psychological science and assess how they place within his account.

Doris's account of the exercise of morally responsible agency is dialogic, but largely focuses on the exercise from one side, the perspective of the doer. What about the other side, the observer or judge? How do people go about making judgments about *others* relevant to morally responsible agency? First, surveys of people across the globe over the last decade allow us to be more certain about what people explicitly value. There is solid evidence that caring and compassion are broadly valued, whereas harm and exploitation are inconsistent with most people's values (e.g., Haidt 2007; Graham et al. 2011). This suggests that in the aggregate, people should not be "victim blamers"; they should attribute blame and responsibility so that harm-doers do not get off the hook, and vulnerable people who have been harmed are protected. In our own vignette studies, this is largely how participants make judgments: People (who were not explicitly labeled "perpetrators") who robbed and sexually assaulted were attributed more responsibility and blame than those who were robbed or assaulted (who were not explicitly labeled "victims"), and people higher in caring values rated explicitly labeled victims more injured and wounded (Niemi & Young 2016). These findings are consistent with findings from moral psychology that demonstrate people's general aversion to directly harming others and their weighting of information about kindness and compassion in person perception (e.g., Greene et al. 2001; Goodwin 2015; Miller et al. 2014). Recent neuroscientific work links moral condemnation of harm to normally functioning emotional processing (e.g., Crockett et al. 2010; Greene et al. 2001; Koenigs et al. 2007; Park et al. 2016; Perkins et al. 2013). Taken together, these findings indicate that equating agency with the term *negotiation* doesn't fit with how people go about moral judgment in cases of direct inducement of bodily harm. In these cases, agency isn't negotiated; it probably never gets close to the negotiation table because most people's biology supports values that reflect concern about bodies as protected from painful imposition.

In an approach complementing Doris's "ecumenical pluralism" (Doris 2015b, p. 186) with respect to agency, continuity, and identity, these massive survey efforts took a moral pluralist approach, going beyond the values of WEIRD participants (people from Western, educated, industrialized, rich, and democratic [WEIRD] societies; Henrich et al. 2010). Findings revealed not only broad shared valuation of care, but also variability in people's endorsement of statements reflecting the values of loyalty, obedience to authority, and concern about purity (*binding values*; Graham et al. 2009; 2011). Strikingly, some people explicitly rank concern about binding values equivalently

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

to concern about "doing no harm," whereas others seem offended by the very idea of such a prospect. People higher in binding values tend to also endorse higher levels of religiosity and political conservatism. We have found that the more people endorse the binding values of loyalty, obedience to authority, and concern about purity (controlling for gender, politics, and religiosity), the more they appear like "victim blamers" – they are more likely to attribute blame and responsibility to victims, say a change in the victim's actions would have made a difference to the outcome, rate victims as contaminated and tainted, and generate fewer counterfactual statements about perpetrator behavior when asked "how could the outcome have been different" (Niemi & Young 2016).

These findings indicate that, in addition to amending Doris's *valuational* theory of moral agency to account for the role of the body and more broadly shared valuation of compassionate caring (i.e., that biologically based aversion to harm allows for some reasonable predictions about action and moral judgment), the proposition that values are central motivators of action is underspecified in another way: Modern culture may be unified about caring, but it's not unified about loyalty, obedience, or purity. And explicit endorsement of binding values is reliably related to how people attribute responsibility and blame across the moral dyad of agent and patient. That is, the extent to which people value obligations at a more abstract level related to loyalty, obedience, and purity relates to how much they factor the contributions of *affected* individuals – moral patients – into their moral judgments. These judgments of responsibility, blame, and contamination have the potential to be consequential to individuals' well-being and personal freedom.

However, consistent with Doris's account, people's judgments were still also influenced by factors outside their awareness. We experimentally manipulated linguistic focus on agents versus patients in vignettes involving sexual assault by placing the perpetrator (agent) in the subject position in the majority of sentences for half the participants, and the victim (patient) in the subject position for the other half. When people focused on victims (patients), they attributed them more responsibility and blame compared to when they focused on perpetrators (agents) – this implicit influence factored into ratings of responsibility and blame in addition to binding values (Niemi & Young 2016). These findings may be taken as some evidence that *judgment* of morally responsible agency across the agent-patient dyad can unfold similarly to how Doris proposes moral agency unfolds from the first-person perspective: as an exercise of values penetrated by implicit influences.

What can psychological science say about how values might relate to perception of the self? In the last chapter of the book, Doris expands on the notion of the *socially contingent self*, crucial to his collaborativist view of moral agency. To do so, he shifts from how individuals are constrained even when they feel their most *able*, to a complementary and illuminating theme: how the severing of meaningful social ties apparently leaves individuals feeling completely *disabled*. In a striking passage, Doris describes how the last-surviving members of the Crow tribe, subjected to cultural annihilation, subsequently reported existential emptiness, as though they had predeceased their bodies. Doris contends that cultural devastation experienced by members of the tribe led to a specific kind of intra-psychological change: rupture in the sense of continuity of the self, as though they were "no longer the same person" (Doris 2015b, p. 183).

Do people really endorse disruptions of personal continuity like this? Indeed they do. Trauma-related cognitions including beliefs about a *foreshortened future* align with the self-rupture Doris describes, e.g., "My life has been destroyed," "I feel like I don't know myself anymore," "I've lost my soul forever," "I feel dead inside," "My life will never be the same again" (Ehlers & Clark 2000; Foa et al. 1999; Nizzi et al. 2012; Nizzi & Niemi, in preparation); as does the experience of *depersonalization*: a feeling of being "unreal" or "detached from oneself" (Yehuda et al. 2015). When these beliefs and experiences can occur in the context of posttraumatic stress disorder, the associated experience of dissociation, or "shutting down," involves inhibition of emotion processing areas in the brain, including the amygdala (Yehuda et al. 2015). The "checking out" response can be contrasted with the (often coexisting, trading-off) response to trauma involving hyperarousal and emotional outbursts (Yehuda et al. 2015).

Most people – estimates are around 50% to 70% of the population (Kessler et al. 1995; Yehuda et al. 2015) – have experienced a traumatic event, such as facing the threat of death, attack, molestation, rape, surviving or witnessing a horrible accident, experiencing combat. The great majority don't develop disabling PTSD (Yehuda et al. 2015), and purportedly don't experience a rupture in sense of self. Doris's theory makes important novel predictions relevant to traumatic experience. First, the more that a person's traumatic event involved profound cultural-level disturbances, or that a traumatized person is prevented from expressing his or her values as a member of a group, the more he or she should report self-discontinuity, as indexed by endorsement of "shutting down" experiences, associated trauma-related cognitions, and inhibited emotion: depersonalization, dissociation, and reports of a sense of a foreshortened future; and not hyperarousal. Furthermore, Doris's theory suggests that remediation of symptoms will come about through a collaborative conversation about values, a position that proposes an intertwining of philosophy and clinical psychology, and one that we support. Finally, it suggests unsettling effects on moral judgment of harm associated with traumatic experience. Specifically, harm to self and others may be judged as more acceptable to the extent that trauma causes "shut down" of emotional processing and an associated rupture in the sense of self, as though one has "predeceased the body." This suggests a mechanism for inter- to intra-group spread of violence: Targeted cultural annihilation may breed callousness broadly (not just retaliatory rage) because targeted, traumatized individuals experience affective shutdown that allows them to more easily harm close others, negatively affecting intra-group relations.

Future research consistent with Doris's pluralist account of moral agency has the potential to link thinkers across the disciplines of philosophy, psychology, and neuroscience. Ongoing investigations indicate that appropriate tools for these investigations will include measures that tap people's explicit endorsements of moral values (e.g., Clifford et al. 2015; Graham et al. 2011); sense of the self as continuous (self-discontinuity scale: Nizzi & Niemi, in preparation; Nizzi et al. 2012); symptoms of avoidance and numbing, i.e., "shutting down" apart from hyperarousal after trauma (Blake et al. 1995); suicidality; as well as measures of neural activity and physiological markers of arousal (e.g., fMRI, EEG, EMG), implicit cognition, and emotional processing.

Doris's account acknowledges, in detail, deep constraints on human freedom. Happily, this theory of constraints has the potential to inspire much creative work, and to engender rich scientific and philosophic questioning about whether and how people exercise moral agency, and about the nature of the self.

Another rescue mission: Does it make sense?

doi:10.1017/S0140525X17000796, e56

Piotr M. Patrzyk

Faculty of Business and Economics, University of Lausanne, Quartier UNIL-Dorigny, Internef, CH-1015 Lausanne, Switzerland.
piotr.patrzyk@unil.ch

Abstract: Two misguided ideas dominate philosophical thinking on moral responsibility: (1) the idea that it obviously exists, and (2) the idea that even if it does not, it is nevertheless needed for the society to function properly. In his book, Doris (2015b) discusses the first illusion, while uncritically

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

accepting the second. In this commentary, I question the utility of such endeavors.

The problem motivating Doris's work is that conscious agency, thought to be required for responsibility attribution, might not be frequently exercised. He finds this morally problematic, as he would like to be able to hold people responsible for what they do. Thus, the goal of his book is to change the definition of agency, such that people can be justifiably deemed responsible more often.

Doris does a good job in providing a *descriptively* accurate picture of how humans attribute exercise of agency. If the goal of his book was to *explain* the pattern of responsibility attribution, its value would be beyond any doubt. Unfortunately, Doris (2015b) is not primarily interested in the descriptive, but in the *normative* question, as he wants to *justify* the practices of responsibility attribution (pp. x, 14). The underlying logic of his book can be summarized as follows (pp. x, 33–34): (1) We often want people to be held morally responsible, (2) if we want people to be held morally responsible, we need a theory justifying such attributions, therefore (3) we need to devise it. Given this goal, the tone of Doris's work is far from neutral. He depicts his attempt not as an investigation of compatibility between the knowledge about human mind and the notion of moral responsibility, but as a "rescue mission" that is supposed to protect the allegedly valuable notion of agency and responsibility from the threat of skepticism. He bravely "resists pessimism" (p. x), "worries" about skepticism (p. 135), but does not "succumb to skeptical panic" (p. 160), such that he no longer feels "skeptical anxiety" (p. 33).

In this commentary I illustrate why such an attitude might be deeply problematic. Determining the degree of agency and responsibility always carries a moralistic flavor (Waller 2011). Can I take credit for my positive acts? Can others blame me for my negative acts? Am I allowed to avenge those who did some wrong to me? When these questions are asked, the responder's reputation is at stake and his or her answers reflect this contingency. Hence, in addition to the intuitive appeal of agency attribution (see Wegner 2002), there is also a moralistic appeal of agency attribution – proclaiming some attribution of responsibility carries evaluative weight, and agents are interested in putting forward their interpretation of reality.

Limits imposed by believability notwithstanding, a general pattern of how responsibility is assigned in social interactions is depicted in Table 1.

Due to the importance of reputation in human groups, it is in an individual's interest to claim responsibility for positive actions and deny responsibility for negative ones (see Alexander 1987). The opposite is true for the actions of others – humans are readily willing to assign responsibility for negative outcomes, as it justifies punitive sentiments toward them, but are not so concerned in the case of positive outcomes (Alicke 2000; Clark et al. 2014). As the human reasoning system is adapted to self-interestedly defend desired conclusions (Mercier & Sperber 2011), the responsibility-related language of excuses and justifications can be seen as a tool for bargaining over the meaning of particular actions (see Bandura 1990; Scott & Lyman 1968).

It is a truism that beliefs and behavior form a feedback system: Beliefs can inform behavior and behavior can inform beliefs. In research on human excuse-making this point is quite clear: It might be both the case that the pattern of excuses people believe in facilitates immoral behavior (Sykes & Matza 1957), but it might also be the case that people form excuses post hoc as they observe and interpret their own behavior (Maruna & Copes 2005; Shalvi et al. 2015).

Table 1 (Patrzyk). *Pattern of responsibility attribution.*

	Actor	
Outcome	Self	Others
Good	Responsible	Not responsible
Bad	Not responsible	Responsible

Experiences and interpretations of single actions contribute over one's lifetime to the formation of generalized beliefs. Just like responses to particular events, these generalized beliefs and philosophical positions should also be seen as a consequence of human predilections (see Cushman & Greene 2012). Topics of philosophical debates are not arbitrary, but reflect the disputants' interests.

If there is individual variability in these beliefs, it might be the case that already formed beliefs influence one's behavior (i.e., they become a self-fulfilling prophecy), but it is also true that individual experiences must have led to their formation. Interestingly, in research on lay notions of free will and moral responsibility, this relationship tends to be treated as unidirectional, that is, the question tends to be framed *not as* "what circumstances foster belief in free will versus determinism," but as "how belief in free will versus determinism influences individual behavior" (e.g., Baumeister et al. 2009; Feldman et al. 2016; Vohs & Schooler 2008; but see Clark et al. 2014).

Such a focus is strange; in situations where human interests are at stake it can be expected that they contribute to the formation of generalized beliefs and moral views (e.g., Deffains et al. 2016). While not denying the importance of discussion of how conscious acceptance of given beliefs influences subsequent decision-making, I would like to focus on a more fundamental question that considers the determinants of belief in agency and responsibility in the first place and interpret Doris's contribution in this light. Why then do humans believe in responsibility?

Agents are usually distinguished from other objects by their ability to self-generate their actions. This intuitive distinction works reasonably well, but can sometimes clash with another lay belief: that agents can, just like objects, be influenced by external factors. Such a belief is usually invoked for denying responsibility and excusing one's own wrongdoing as determined (Bandura 1990; Sykes & Matza 1957).

When one realizes that humans can be influenced, then the "ideal" exercise of agency (i.e., libertarian self-determination of one's actions) becomes untenable (see G. Strawson 1994). As people want to retain both intuitive theory of agency and the possibility of being influenced, they "adjust" their concept of agency such that it does not need to be fully independent from influencing factors (see Monroe & Malle 2010).

That leads to a practical problem. When one wants to assault someone else with the charge of responsibility for immoral behavior, one needs to concoct a justification for this attribution. One is perfectly aware that there exists a long causal chain of past events that has eventually led to a given behavior (even if they only influenced, rather than determined it), but assigns exclusive moral responsibility to the final action executor only (unless *I* am the wrongdoer, in which case the importance of external factors is obvious). That is to say, even if one is aware that the "ideal" agency was not exercised, one acts as if it was, "forgetting" about the prior history.

As this is objectively unfair, creating a situation in which blame judgments depend on factors that cannot be controlled by the actor (i.e., it forms so-called moral luck problems), it leads to the challenge of how to justify the entire practice. Philosophers seem especially reluctant to honestly admit that the only reason for such a blame assignment is to satisfy people's thirst for revenge, so they engage in the task of finding different justifications (but see Waller 2011). The most obvious one, however, a utilitarian appeal to convenience and societal benefits following from revenge, is also usually thought to be insufficient (e.g., P. Strawson 1962; Smilansky 2000). As the argument goes, there must be "deeper" reasons justifying current practices.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

One prevalent proposal consists in denying the existence of determinants. Past events might have happened, but the final actor has the power to counteract them, as he or she exercises conscious agency. Doris (2015b) is dissatisfied with that approach primarily because he would like to accuse some as responsible in cases where conscious agency did not take place and excuse some others as not responsible even if conscious agency took place (e.g., pp. 24, 128). Thus, he approaches the problem in a different way. Instead of outright denying the existence of determinants or calling for a wishful disbelief in them (see e.g., Dennett 1984; Smilansky 2000), he accepts that conscious agency is not usually exercised, but argues that this fact might or might not matter for responsibility attribution (p. 11).

Doris (2015b) rejects the normative importance of conscious thoughts and proposes "pluralism about agency and responsibility," which is a convenient euphemism for arbitrariness. Criteria for responsibility attribution may vary across different circumstances in a way that ultimately follows lay intuitions (pp. 171–75). Such a proposal is nothing new; Strawson's (1962) classic treatment posited finding a justification for existing institutions in the intuitive appeal of our retributive instincts. Doris moves the problem one level up by positing that responsibility should be determined by evaluating if the offered explanation agrees with the values present in actor's lifetime and if it is approved by others. Naturally, if determinism is true, all of this is equally independent from the actor in the same way atomic actions are. Because of that, he needs to appeal to intuitions over when determination should imply responsibility and when it should not.

Let me give an example of what divergent responsibility judgments Doris (2015b) seeks to justify: When one claims responsibility for choosing the right numbers in a lottery, we ridicule it (p. 134), but when someone has worked hard and claims responsibility for his or her life successes, we often uncritically grant it. Such an intuition is to be expected if one's source of money is work, rather than gambling, but it does not change the fact that this is arbitrary and unjustified, as both agents benefit from mechanisms beyond their control, be they genetic, environmental, or random. What is then the value of a theory that, instead of realizing unfairness of some attributions, seeks to "legalize" them?

I am not so much concerned about *how* Doris proceeds in "rescuing" agency and responsibility; in the end, if the work starts with an assumption that the existence of agency is good and the lack of agency is bad, it should not come as a surprise that Doris found agency to exist (in the cases where it conforms to his intuitions). What I find more problematic is that Doris (2015b) does not make a compelling case about *why* we actually need to devise normative notions of agency in the first place. The only reason for believing in agency and holding others responsible is that it is supposed to be useful (pp. x, 136). But this point, presumably assumed to be so clear that is does not require any explanation, is in fact not clear at all. Doris claims that failure to justify the existence of agency would lead to pessimism (p. x), that he does not like approaches leading to the disintegration of the mind and feels responsible for restoring the sense of personhood and agency (p. 3), or offers general remarks that bad theorizing can lead to "morally reprehensible" consequences (p. 9). But what *exactly* would happen if we stopped believing in agency and responsibility?

The answer to this question is not obvious as there are divergent intuitions over what, if anything, would change if the belief in moral responsibility was abandoned (e.g., Caruso 2014; Waller 2011). Unless anti-skeptics have some sound conception of consequences of finding or not finding a justification for moral responsibility, the motivation behind these contributions needs to be taken with caution. If the proposed explanation boils down to a conjecture that agency exists because one wants it to exist, one will always find a reason why it is so; Doris's work is neither the first nor the last attempt of this kind. Persistence of anti-skeptics in finding solutions to their own problems raises the question of what inspires these "rescue missions." Taking into consideration the strategic context in which the social construction of responsibility takes place, I am afraid that there is no other plausible explanation for these attempts other than the need to devise a theory allowing them to justify their retributive instincts and take credit for what they do not deserve (see also Miles 2015).

The tangled web of agency

doi:10.1017/S0140525X17001170, e57

Alain Pe-Curto,[a] Julien A. Deonna,[b] and David Sander[b]

[a]*Department of Philosophy, College of Arts and Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3125;* [b]*Swiss Center for Affective Sciences and Department of Philosophy at the Faculty of Humanities, Université de Genève, 1202 Geneva, Switzerland.*

alain.pecurto@unc.edu http://philosophy.unc.edu/people/alain-pe-curto/
julien.deonna@unige.ch
http://www.affective-sciences.org/en/home/cisa/members/julien-deonna/
david.sander@unige.ch
http://www.affective-sciences.org/en/home/cisa/members/david-sander/

Abstract: We characterize Doris's anti-reflectivist, collaborativist, valuational theory along two dimensions. The first dimension is social *entanglement*, according to which cognition, agency, and selves are socially embedded. The second dimension is *disentanglement*, the valuational element of the theory that licenses the anchoring of agency and responsibility in distinct actors. We then present an issue for the account: the *problem of bad company*.

Talking to Our Selves ends on the note "Afterwards." The note encapsulates the picture developed in the book, based on abundant references to fascinating empirical findings with implications for our status as morally responsible agents and selves. As he concludes, Doris (2015b, p. 199) enjoins us not to see ourselves as "little gods with big brains," but as "social animals that endlessly explain themselves to one another" in ways that resort much less to "the rarefied sort of [cognition] called Reason" (p. 199) than philosophers of a Kantian or Frankfurtian bent have suggested (Doris 2015b, pp. 7, 18–19; Shoemaker 2015). The empirical evidence shows that "the mental is rather a mess" (Doris 2015b, pp. 3–5); in particular, messier than reflectivists, who emphasize the importance of reason for agency, could tolerate. "Minds, we're, told are uncertain conglomerations of systems and subsystems – 'modules' on one influential family of theories – cobbled together by natural selection with little regard for the people toting them about" (Doris 2015b, p. 3).

On the face of such a "disintegrated" picture of the mind, with competing processes yielding incongruent outputs, one could either become a skeptic regarding agency and conclude that we rarely are *bona fide* agents, or attempt to discount the empirical evidence. Doris elects a third path: He ambitiously presents an account of moral agency and responsibility that is empirically informed: "As a philosopher of morals, I don't think [the picture of mental disintegration is] sufficiently enriching; with human beings, I want to find a there there – a person, rather than a haphazard muddle of cognitive systems" (Doris 2015b, p. 3). His concluding note comes at the end of an important and ingenious piece that builds on an enduring worry (see Doris 2009). This piece is another crucial building block in Doris's project of developing an empirically adequate moral psychology, one that complements his work on character. Like his previous work, it promises to reshape the debate with which it engages.

The author's starting point is the "individualist," "reflectivist" notion that *"the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do"* by the individual, and its corollary, that *"the exercise of human agency requires accurate reflection"* by the individual (précis, sect. 3, para. 1, emphasis in the original). Doris lays out

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

the reasons to believe that individualist reflectivism fails and proposes an anti-reflectivist, collaborativist account in its stead.

As he did for character (Doris 2002), the author's objective is to recalibrate the debate. He puts collaborativism center stage, but does not fully exclude individual, reason-based agency. We may sometimes act in ways that approximate reflectivist, individualist agency, but it is neither the most common nor the optimal form of agency (Doris 2015b, pp. 22–23, 103 sqq.). Doris's account includes a valuational component as well: "it locates the exercise of agency in the expression of a person's values" (p. x). Through this valuational component, the account has implications for moral responsibility and for the (synchronic and diachronic) unity of our selves. The identity and continuity of a person is tied to her subjective values. Moreover, "attribution of agency and responsibility may be warranted when a pattern of cognition, rationalization, and behavior emerges, and that pattern is best explained as involving" the expression of her values (p. 164).

We focus here – as does Doris – on the opposition between individualist reflectivism and collaborativist anti-reflectivism (Doris 2015b, pp. 110 sqq.) and concentrate on the collaborativist aspect. We first characterize the account along two dimensions, before turning to what we call the problem of bad company.

The first dimension is entanglement: The solution to the skeptical challenge is that agency and cognition are not fully internal to the agent or, in philosophers' patois, intrinsic affairs. They are, most commonly and when optimal, entangled in a social context. "If reflectivism engenders skeptical difficulty, individualism obscures the best avenue of escape. Ameliorating skepticism about agency will require rejecting *both* reflectivism and individualism" (Doris 2015b, p. 110). So, most common and optimal forms of agency are collaborativist – they are embedded in a "dialogic" relation with others (pp. 146–48). We can resist skepticism if we abandon the idea that we have an instance of agency only when we have a rational agent reflecting self-consciously in a condition of "freedom from [social] influence." Instead, we need to embrace "mutual influence" (p. 148).

Entanglement also applies to selves: We are socially embedded beings, the synchronic and diachronic unity of which is "contingent on" social context (Doris 2015b, pp. 191–92, 196). What is distinctive to the account here is not *contingency*. Rather, it is the idea that a self is and remains itself not solely in virtue of intrinsic considerations. Her identity and continuity is dependent on extrinsic factors located in her social environment (Ch. 8). This claim may seem unsurprising, but we should note that the dependence defended by Doris (2015b) is strong: It is not only that individuals need social context as a "precondition" to develop and sustain their identities. It is also a deeper claim of "causal entanglement" (pp. 191–92). Cognition, agency, continuity, and identity do not simply require social context; they are *characteristically* social (pp. 115, 191–92).

This social nature is key to responding to the skeptical challenge. Through the exchange of *rationalizations* and *social* negotiation of our "biographies" (Doris 2015b, pp. 143 sqq.), we can exercise agency in our lives by behaving in ways expressing our values:

> The social exchange of explanatory and justificatory narratives erects a scaffolding that supports behavior expressing the actor's values. These dialogs effect the exercise of agency in conditions of self-ignorance where direct reflective control may falter. In the right social milieu, agency obtains in spite of – or rather *because* of – self-ignorance. (Doris 2015b, p. 129; on self-ignorance, see p. 21)

Notably, we are often deluded regarding the causes and reasons for our actions and our degree of control over our behavior. But these embellishments, both sustained and limited by appropriate social context, in fact, favor agency (Doris 2015b, pp. 146 sqq.).

The second dimension along which we characterize Doris's account is disentanglement. Indeed, one might wonder if the emphasis on collaborativism does not leave us with agents and selves, the boundaries of which are unclear. The author refrains from claiming we are "extended selves," incorporating (as parts) elements of their social context – even though he confesses a certain temptation to do so (Doris 2015b, pp. 191–92). However, the stronger the evidence for skepticism about individualist reflectivism, the stronger the case for a substitutive account. And as the evidence for such skepticism accumulates, the more Doris's substitutive account needs to emphasize collaborativism, the "best avenue of escape" from skepticism (p. 110).

If agency and persons are socially embedded so profoundly, we could wonder: How are we still *individually* responsible agents and *distinct* selves? The valuational component of the theory plays a key role in this respect. For Doris (2015b), a criterion to identify what is internal to the self is (the expression of) the actor's values (pp. 24–25). He calls behavior that expresses the actor's values "self-directed," and such self-directedness is necessary for agency and responsibility. These values, although socially contingent, individuate selves: "A person has the identity they do partly by virtue of having the values they do: who I am has much to do with what matters to me" (p. 188).

According to our characterization of Doris's theory, the entanglement dimension helps respond to skepticism, while the disentanglement dimension licenses the anchoring of agency and responsibility in distinct actors. Let us now consider a thought experiment and present a potential issue: the problem of bad company.

Imagine Cleo contemplates a life-defining choice – say Cleo is deciding between philosopher and clinical psychologist careers. She has purposely isolated herself for days from all external influences, such as that of her philosopher friends and psychologist parents, who tend to sway her in one direction or another. Cleo wants her career-related behavior to be the result of *reflection* and to originate *from herself*. Cleo is a reflectivist individualist about agency. She wants her choice to be made in a condition of "freedom from influence."

According to collaborativism, however, social dialogue – through explicit or implicit pathways – might help her behave in the manner that expresses her values the most appropriately. Instead of seeing influence as a corrupting force, we could see it as a positive fixture of agency. For example, sociality could mitigate some "goofy influences" (Doris 2015b, p. 64), that is, rationally irrelevant elements that seem to surreptitiously affect one's cognition and behavior. Assume, for the sake of example, that when Cleo listens to jazzy music, she feels drawn to philosophy and, conversely, that a taste of Anthony's freshly baked cookies tends to make her desire to help people by becoming a therapist (see Ch. 3; the example is inspired by Isen & Levin 1972). Collaborativism seemingly advises Cleo and us: Don't worry too much about the relative lack of conscious, individual control on your behavior – self-ignorance and mutual influence is right for you! As Doris says, "ignorance is the life-blood of agency," "the social formation of biography facilitates agency," "relationships help people express their values in their lives, as they do in the right sort of friendships, romantic involvements, and institutions" (pp. 129, 146, 148).

Let us grant that, with mutual influence, we are better placed to express our values and that we should embrace our "weathervane-like" natures, marked by less conscious, individual control than reflectivist individualism assumes. Doris assures us that the winds are mostly propitious. They are fairly steady as well, bringing stability. Doris (2015b) speaks, for instance, of the social "scaffolding" or "mast" that our socially negotiated biographies constitute (pp. 129, 144). They support individual responsible agency and promote morality, thanks in part to "self-bullshitting," (Shoemaker 2015) *pace* reflectivism (Doris 2015b, pp. 143–45, Ch. 6; see also Alfano 2013a).

Yet, the troubling discovery that we are often clueless as to the determinants of our behavior, that we are like weathervanes, lingers. Thus, we could turn Doris's epistemological argument against reflectivism on its head and ask: Are we sure that the winds are at our backs? We do not need to search long before we find examples of human groups leading individuals to act in atrocious ways: socially sustained instances of slavery, genocide,

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

and forms of nasty implicit and explicit "-isms" are unfortunately countless. Doris responds that, indeed, the type of social environment with which we surround ourselves matters – it must be "the *right* social milieu" (our emphasis). But how is Cleo, as an individual, to tell whether philosophers or psychologists are bad or good company? And can we not, as a community, err in a "self-ignorant" manner? Bad company appears abundant enough, but, even if one assumed that sociality most often supports individual responsible agency and bolsters morality, what if, sometimes, *collaborativism* sounds too much like *collaborationism*?

Doris (2015b) does note that "sociality . . . facilitated resistance to the Nazis" (p. 119) and that some moral emotions – often seen as social emotions – promote individual morality (pp. 121–22; on, e.g., shame, see Calhoun 2004; Williams 1993). But showing that there is good company or even that bad company is rare does not suffice. We need to *rule out* bad company, which is an analog to the "more moral than purely epistemic" defeaters (précis, sect. 2, para. 8) that plague individualist reflectivism (Doris 2015b, pp. 64 sqq.). If we are so entangled in our milieu for cognition, agency, and our unity as selves, we appear badly placed to tell whether we should embrace or resist its influence. Should we be in bad company, we might need *reflective*, *individualist* humans sufficiently in touch with their values, and so able to disentangle themselves from social influence. If we may be so bold to suggest it, we might need humans of character.

ACKNOWLEDGMENTS

We owe the title of this commentary to Otto Bruun (and Shakespeare). We thank him for it and for his insightful and indispensable comments as well. We are also very grateful to the members of the Swiss Center for Affective Sciences and the Department of Philosophy of the University of Geneva for enlightening exchanges, in particular Elena Cagnoli Fiecconi, Mary Carman, Florian Cova, and Steve Humbert-Droz. We thank the Swiss National Science Foundation, the Société académique de Genève, and the Dean's Office of the Faculty of Humanities of the University of Geneva for their support.

Negotiating responsibility

doi:10.1017/S0140525X17001169, e58

Tamler Sommers

Philosophy Department, University of Houston, Houston, TX 77204.
tamlers@gmail.com
http://www.uh.edu/class/philosophy/people/somers/

Abstract: I argue that John Doris should apply his dialogic or collaborationist approach to agency more fully to questions of moral responsibility. To do so, he must discard his form of pluralism that aims to accommodate a variety of theoretical approaches to responsibility in favor of a pluralism that rejects theorizing about responsibility altogether.

In part I of *Talking to Our Selves*, John Doris marshals impressive evidence to challenge "reflectivism," the view that human agency consists in judgment and behavior ordered by self-conscious, accurate, reflection about what to think and do (Doris 2015b, p. 17). In Part II, Doris outlines an alternative account – the "dialogic" or "collaborationist" view – that holds that (a) agency involves the expression of human values, and (b) we express and discover our values through an ongoing social process: via conversation, rationalizing, apologizing, feeling guilty, grateful, angry, forgiving, and so forth. "Human beings living in groups," Doris writes, "shape their lives not as isolated reflectors, but as participants in an ongoing negotiation – a negotiation that simultaneously constrains and expresses who they are" (Doris 2015b, p. 148). Agents, he concludes, are negotiations.

Others are better qualified to comment on the evidence Doris mounts against reflectivism. I found Doris's criticisms compelling, and don't find the conclusion nearly as depressing as many of my fellow philosophers. Indeed, the dialogic account seems both more attractive and better coheres with my own experience. My criticisms will focus on how Doris applies the dialogic account to questions about moral responsibility.

For Doris, responsibility involves agency, and because agency involves expressing values, morally responsible behavior is behavior that expresses the agent's values. Because we often don't know whether an act expresses an agent's values, we can be agnostic about the agent's responsibility in those cases, or try to learn over time where the agent stands. Doris notes as well that many actions (infidelity, for example) often express more than one conflicting value. He argues that we bear greater responsibility for wholehearted actions than we do for our ambivalent or conflicted actions. He even concedes that in certain cases reflective deliberation might also be necessary for agency. In those cases reflective deliberation might be required for responsibility as well. Doris calls himself a pluralist about both agency and responsibility, though he clearly favors the dialogic view.

Philosophically minded critics are likely to complain that Doris's account is too indeterminate, pluralistic, and messy. My complaint is from the other direction: I don't think Doris is being indeterminate, pluralistic, and messy enough. Doris, I'll argue, should steer his dialogic approach deeper into the realm of responsibility judgments. To do so, however, he must discard his form of pluralism that aims to accommodate a variety of theoretical approaches to responsibility, in favor of a pluralism that rejects theorizing about responsibility altogether.

Doris takes a step in the anti-theoretical direction by embracing a broadly Strawsonian view of responsibility that associates responsibility judgments with the range of interpersonal reactive attitudes, such as resentment, gratitude, forgiveness, love, guilt, and pride. But Doris pulls back when he concedes that the attitudes and practices "don't carry all the normativity that might be desired." And he gives up the game entirely when he claims that the way to establish normative authority "is by reference to theory." Attributions of responsibility, Doris writes, only have normative authority when they can be seen "to have a rationale sufficiently compelling to serve as a justifiable guide to thought and action." So even though Doris understands responsibility judgments via the reactive attitudes, he accepts the challenge "to identify compelling theoretical grounds for when and what reactive attitudes are appropriate."

Strawson himself doesn't make these concessions. He insists that the general framework of reactive attitudes are simply given with the fact of human society. Consequently, the framework "neither calls for, nor permits, external "rational" justification' (Strawson 1962). Strawson offers no necessary and sufficient conditions for appropriate instances of the reactive attitudes; he believes that our actual human attitudes and practices get the final say. According to Strawson, requiring theoretical grounds to justify the reactive attitudes "overintellectualizes the facts" about responsibility. So on this issue (and only this issue), Doris overintellectualizes the facts. And this is unfortunate, because Doris's account of agency is very much in the spirit of a truly Strawsonian approach to responsibility.

To see what I mean, consider cases in which an agent's behavior doesn't express his values at all. One example Doris discusses involves a father who accidentally leaves his sleeping child in the car on a hot day, resulting in the child's death. Because acts like these don't involve the expression of values, Doris is inclined (with certain caveats) "to eschew attributions of moral responsibility for slips." My inclination, in contrast, is to resist making generalized responsibility judgments about such cases at all. I don't think it's the philosopher's business to cast judgment on the father's responsibility. Rather, it's up to him, his family, and their community to arrive at this determination.

To make the issue more concrete, consider a similar kind of case from the recent film *Manchester by the Sea* (mild spoilers to follow; Lonergan 2016). At the beginning of the movie the protagonist, Lee Chandler (played by Casey Affleck), is clearly burdened

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

with tremendous pain and guilt, but we aren't sure for what. We learn the answer in the middle of the film. One night, eight years earlier, Lee had been drinking with a group of friends at his house. At 2 a.m. his wife Randi sent his friends home for being too loud, concerned that they would wake the kids. Too wound up to sleep, and too drunk to drive, Lee decides to walk to the nearest 24-hour grocery store to get a six-pack of beer. His wife has been sick, and the heating system dries out her sinuses, so their wood fire was the only source of heat. Lee leaves the fire going to warm the house, but forgets to put the screen on the fireplace. A log rolls out. When he returns from the store, the house is on fire. His wife is rescued but their three children burn to death.

On Doris's account, Lee would not be morally responsible for the deaths of his children. His act was a terrible accident that did not express his values in any way. Lee himself, we learn, emphatically rejects this verdict. During his confession at the police station, Lee is shocked and angered that the officers won't charge him with a crime. He wants to be held responsible, he wants to be punished. He think it's *right* for him to be held responsible and punished. His wife Randi is similarly disinclined to let Lee off the hook, at least in the weeks following the incident. And the same seems to be true for most of the people in the small town of Manchester, where they live.

A few things to note about the case. First, the relevant facts aren't in dispute. Nobody thinks that the act secretly reflected Lee's values. Nobody thinks it was the result of deliberation or reflection. They're aware of the facts, but they hold him morally responsible anyway, especially Lee himself. Second, Lee would take absolutely no consolation from a "not responsible" judgment issued by a theory of responsibility. It would mean precisely nothing to him. The only verdicts that matter to Lee are the verdicts of his wife, family, friends, and neighbors. And that seems right in this case. Why should philosophers or theories get to determine Lee's moral responsibility for this tragedy? Why shouldn't we leave it up to Lee, Randi, and the people who are involved in the situation to arrive at – to negotiate – their own responsibility judgments based on what they regard as relevant to the situation?

Agency is a negotiation, Doris claims. Responsibility is often a negotiation too, but not one between compatibilists, incompatibilists, and other responsibility theorists. It's a negotiation between the participants, the people affected by the action. Accepting this means accepting indeterminacy about responsibility, accepting that that there are multiple plausible outcomes even under similar or almost identical circumstances. Perhaps if Lee and his wife had different temperaments, or if they lived in a different town, or if they were more religious, or less religious, they might arrive at a different verdict. Maybe Randi could find a way to forgive Lee, and Lee might even eventually forgive himself. That could be a plausible outcome of negotiation too. Indeed, perhaps the most tragic element of *Manchester by the Sea* isn't that Lee is held morally responsible for an accidental slip. The tragedy is that he is so grief stricken that he *shuts himself off from dialogue*. He shuts himself off from negotiation. In a haunting scene eight years after the tragedy, Randi asks Lee to have lunch. She wants to apologize for some of the things she said to him after the incident. But Lee is too broken at this point to continue the dialogue.

Philosophers at this point might be tempted to pounce. *You see*? *If only Lee had accepted our theories, he would have understood that he is not technically blameworthy for what happened.* But let's be honest. No theory would have a prayer of influencing him on this, nor should it. For Lee to be moved at all by such considerations, they would have to come from a friend, a family member, or perhaps a priest, not from philosophical busybodies and their theories. For Lee to arrive at a different verdict, it would come through dialogue and negotiation with participants

I suspect that Doris will have a great deal of sympathy with the view that I've sketched here. Indeed, he may claim that it's consistent with his *pluralism* and *variantism* about responsibility (see Doris 2015b, pp. 171–77). But as I suggested above, Doris's brand of pluralism and variantism is designed to accommodate the normative authority of multiple theories, rather than deny their authority entirely. A thoroughly dialogic approach would resist the urge to offer an account with "all the normative weight that might be desired." The participants in relationships must provide the normative authority, or at least most of it.

Does this mean that philosophers have no role to play in evaluating the appropriateness of moral responsibility judgments? Not at all. There's great value in identifying and describing common trends embedded in our responsibility practices and their associated attitudes. Philosophers (and responsibility theorists from other fields) can also call attention to the mistaken empirical assumptions that people often have concerning agency and control. This is what Doris does so well in part I of *Talking to Our Selves.* This would not count as busybody theorizing if the mistaken empirical assumptions lead people to make responsibility judgments that are false *by their own lights.* What philosophers should resist is the temptation to theorize about how it is rational or appropriate to feel about this new information. Indeed, Doris's dialogic account of agency points the way toward a principled source of resistance.

Agency enhancement and social psychology

doi:10.1017/S0140525X17000802, e59

Matthew Taylor
Philosophy Department, Florida State University, Tallahassee, FL 32306.
Mattaylo87@gmail.com **matthew-c-taylor.com**

Abstract: This commentary has two aims. First, I raise a practical challenge for accounts of responsible agency: Provide empirically informed strategies for enhancing responsible agency so that actors can become more resistant to the influence of defeaters. Second, I offer an initial sketch of a solution to this practical challenge. My solution is supported by empirical evidence suggesting that responsible agency can be enhanced via self-regulatory strategies (expertise and implementation intentions).

In *Talking to Our Selves*, John Doris claims that attributions of responsible agency may be warranted when the best explanation for certain patterns of behavior across iterated conditions involve the expression of the actor's values. In some cases, however, attributions of responsibility will be unwarranted because we cannot rule out the presence defeaters. For any actor A, behavior (or cognition) X, and practically significant causal factor(s) F, defeaters obtain when the causes of A's X are such that, had A become aware of F, A would not recognize F as reason(s) justifying A's X. In cases where defeaters obtain, exercises of responsible agency do not obtain. If we cannot rule out the presence of defeaters, then our attributions of responsibility are unwarranted.

I want to expand Doris's discussion by further focusing on cases where we are unable to rule out the presence of defeaters. In particular, I am interested in the practical issue of enhancing our responsible agency so that defeaters no longer have a significant impact on behavior and cognition. I maintain that any practically relevant view of responsible agency should have something to say about circumventing the influence of defeaters. Doris does provide an account of how responsible agency may be realized via an extended process of interpersonal negotiation wherein people identify and develop their values. However, it is not entirely clear that these dialogic processes will circumvent the influence of defeaters. Even if I identify and develop my values, defeaters may still bypass those values. In that case, what is needed is an account of how we might become more resistant to the influence of defeaters. Ideally, this account would employ empirically informed self-regulative strategies that enjoy wide

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

applicability. In what follows, I provide an initial sketch of such an account.

Imagine that we have observed a subject, Amy, across various iterated conditions. Suppose that Amy values helping others, but we are unable to infer from her behavioral pattern that defeaters do not obtain. She does not engage in helping behavior in the following iterated conditions: when she is in a neutral mood (O'Malley & Andrews 1983), when she is in the presence of passive bystanders (Darley & Latané 1968), when her willpower has been depleted (DeWall et al. 2008), and when the potential recipients of her assistance are members of a specific race (Stepanikova et al. 2011). However, Amy does help when her body posture has been mimicked by an experimental confederate (van Baaren et al. 2004), and when she is in various other defeater-suspect conditions. Given Amy's extended behavioral pattern, it seems plausible to think that we are unable to rule out the presence of defeaters. She would not recognize mimicry effects, group effects, ego-depletion effects, mood effects, and implicit bias as influences that justify her conduct. However, it may be the case that her behavior was influenced by all of these rationally and ethically arbitrary influences.

Imagine that Amy learns that there are various defeater influences on her conduct. She is deeply troubled by this information. Amy thinks of herself as someone who explicitly devalues racist categories. She also thinks of herself as someone who is not easily influenced by numerous other strange factors. Amy decides to visit a therapist for practical advice about how to enhance her responsible agency so that defeaters no longer have a significant impact on her behavior. Is there any practical advice that can be given to Amy? Here's one possible response. We reply that, unfortunately, she cannot enhance her agency so that defeaters no longer have a significant impact on her behavior and cognition. We might support this claim with the following observation: Behavior is overwhelmingly governed by automatic processing, and our ability to override our automatic processing is severely limited (Bargh & Chartrand 1999). Some social psychologists estimate that our ability to override automatic processing is limited to about 5% of the time (Baumeister et al. 1998; Bargh and Chartrand 1999). In light of this evidence, someone might respond to Amy in the following way: Most of us often lack the ability to enhance ourselves in such a way that defeaters no longer have a significant impact on behavior and cognition.

Amy does not find this response very appealing. Amy isn't just concerned about attributing responsible agency to others, she is just as concerned with enhancing her own agency so that attributions of responsibility apply to her own conduct. She has practical aspirations to circumvent the influence of defeaters across various conditions. Furthermore, I take it that most of us share Amy's practical aspirations: We aspire to actively express our values rather than being passive victims of defeaters. A practically relevant account of responsible agency should have something to say about these important practical aspirations. I maintain that we can provide an initial sketch of such an account. This sketch should provide some optimism for thinking that we can give Amy (and others) widely applicable (and empirically informed) practical advice.

Are there any self-regulatory strategies that are widely applicable with respect to mediating the impact of defeaters? A growing amount of evidence suggests that the development of expertise is a widely applicable self-regulative strategy. Cramer et al. (1988) measured the helping behavior of registered nurses and nursing students in the presence of passive bystanders. An experimental confederate disguised as a maintenance worker loudly fell from a ladder in a nearby room. The helping behavior measured in this experiment involved offering assistance to the confederate. Experimenters found that the nursing students were significantly less likely to help in the presence of bystanders. However, registered nurses were just as likely to help in the presence of passive bystanders as control subjects who were alone.

Cramer and colleagues explain that "the nurses' professional training and experience led to a consistent level of emergency responding whether or not a bystander was present" (Cramer et al. 1988, p. 1142). In addition to this study on group effects, practice and expertise has been found to mediate the influence of implicit bias (Plant & Peruche 2005; Plant et al. 2005), mood effects (Forgas et al. 2008), ego depletion (Hui et al. 2009; Muraven et al.1999), and various other defeaters (Morewedge et al. 2015). Crucially, some of the evidence on expertise suggests that practiced subjects need not detect defeaters nor override automatic processing. Instead, the evidence suggests that practice modifies the actor's automatic processing so that it is less sensitive to defeaters prior to encountering those influences (Forgas et al. 2008). Thus, there is a growing amount of evidence suggesting a widely applicable self-regulatory strategy: People should develop expertise in certain domains if they want to consistently express their values.

The empirical evidence in support of enhancing responsible agency is not exhausted by the studies on expertise. Peter Gollwitzer and other social psychologists have found that adopting situation-specific plans can mediate the influence of defeaters. They propose that people adopt implementation intentions, "if-then" plans that specify a specific context and desired response (Gollwitzer 1999). For example, suppose I'm trying to avoid buying beer at a local restaurant. I can adopt the implementation intention: "If I'm handed a beer menu, then I will order water instead!" There have been a large number of studies suggesting that implementation intentions can facilitate goal pursuit across various domains (Gollwitzer & Sheeran 2006).

Mendoza et al. (2010) measured whether implementation intentions mediate the influence of implicit bias in a computer shooting task. The shooting program displayed either a Black or White face containing either a gun or a non-threatening object. The subjects must respond as quickly as possible (less than 630 ms) to each trial by choosing "shoot" or "don't shoot" in response to various images. They were instructed to only shoot at armed criminals. Subjects in the experimental group were told to adopt the implementation intention: "If I see a person, then I will ignore his race!" (Mendoza et al. 2010, p. 515). Subjects in the control group were not given the instruction to adopt implementation intentions. Experimenters found that control subjects were more likely to shoot unarmed Black targets than unarmed Whites. However, experimental subjects were significantly less likely to make performance errors (i.e., selecting "shoot") in response to both unarmed Black and White bystanders. This evidence suggests that people can mediate the influence of implicit bias via the adoption of implementation intentions.

The effectiveness of implementation intentions is not restricted to mediating implicit bias. Implementation intentions have also been found to mediate framing effects (Trötschel & Gollwitzer 2007), mimicry effects (Wieber et al. 2014), ego depletion (Webb & Sheeran 2002), and mood effects (Webb et al. 2012). It is important to point out that the evidence on implementation intentions suggests that subjects can regulate defeaters without needing to override their automatic processing. This is because the 'if' component of the plan modifies the actor's automatic processing prior to encountering the specific circumstance, thereby shifting attention away from potential defeaters and toward goal-relevant situational cues (Gallo et al. 2009). Thus, there is a growing amount of evidence suggesting that implementation intentions can mediate the influence of defeaters.

It should be admitted here that there may be some limits to the self-regulatory strategies described in my initial sketch. Lauren Olin and John Doris have noted some limits on expertise (Olin & Doris 2014). They discuss some studies suggesting that physicians become worse at diagnosing heart sounds after years of practice (Choudhry et al. 2005; Ericsson et al. 2007), and experts are not better than non-experts at making economic predictions (Tetlock 2009). Thus, it seems that practice sometimes results in worse performance, and it may well be that some of these failures involve defeaters.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Doris also claims that implementation intentions may be limited because "countervailing pressures may perturb the goal relevant situation" (Doris 2015b, p. 128). Someone might also worry that implementation intentions may collaborate or produce defeaters in specific circumstances. The increased salience of situational cues contained in the "if" component of one's plan may automatically attract attention even when these cues are not relevant to one's current goal (Parks-Stamm & Gollwitzer 2009). There is evidence that the human capacity for attention is limited (Shiffrin & Schneider 1977), and so implementation intentions may influence us in such a way that we are sometimes unable to attend to relevant contextual cues during the pursuit of multiple goals. Perhaps even some of these failures of attention may be construed as defeaters, where the causes of the actor's behavior (i.e., the 'if' component of the plan together with the situational cue) would not be recognized as reason(s) justifying their behavior.

Admittedly, the strategies suggested here may be limited in various ways. However, the purpose of my discussion was only to present an initial sketch that enjoys some empirical support. I have not attempted to provide a fully developed account of enhancement that can adequately address all of these potential limitations. Obviously, the details of this account need to be developed further so that these limitations are adequately addressed. But I do not see sufficient evidence for thinking that this project is unlikely to produce a compelling response to Amy (and others). In conclusion, accounts of responsible agency face an important practical challenge: Specify conditions and strategies wherein actors can circumvent the influence of defeaters. I hope to have provided an initial sketch of such an account.

Responsibility: Cognitive fragments and collaborative coherence?

doi:10.1017/S0140525X17000814, e60

James S. Uleman,[a] Yael Granot,[b] and Yuki Shimizu[c]

[a]Department of Psychology, New York University, New York, NY 10003; [b]Yale Law School, New Haven, CT 06520; [c]Faculty of Education, Saitama University, Saitama, Saitama Prefecture 338-8570, Japan.

jim.uleman@nyu.edu yael.granot@yale.edu
shimizu@mail.saitama-u.ac.jp jimuleman.com psych.nyu.edu/Uleman
http://en.saitama-u.ac.jp/research/researchers/dr-yuki-shimizu/

Abstract: We describe additional research that expands upon many of Doris's points, focusing on collaboration (Ch. 5), selves, and identity (Ch. 8). We also suggest some elaboration of his treatment of dual process theories (Ch. 3). Finally, we ask whether collaborationist accounts confer logical consistency.

This rich book should be read by psychologists and philosophers alike, because it introduces a wealth of relevant research, ideas, and references. Doris's (2015b) thesis is that judgments of moral responsibility are relatively independent of notions of freedom and determinism (he's a compatibilist), do not depend on accurate self-knowledge, but do depend on social negotiations and social context. This means that often there is no objectively "correct" or single answer to questions of who is responsible; it depends on context. Some effects of context on moral judgments (e.g., nonconscious priming) cannot be explained to others, while other effects can or might be. But this depends on finding common ground with others. We comment on four core ideas – collaboration, dual process theories, self, and identity – and suggest that social coherence in reflections about moral agency may depend on culture. Finally, we ask whether social coherence confers logical coherence.

"Collaboration" (Ch. 5) refers to the thesis that human reasoning, including reasoning about morally responsible agency, is social and negotiated rather than principled and based on mental states. Thus, accurately reading mental states is less relevant than reflectivists contend. Haan made a similar point and contrasted it with Kohlberg's rationalist view of moral development. "[M]oral truth is based on agreements moral agents achieve about their common interest and is not predetermined by rules or principles, that is, truth is to be achieved, not revealed" (Haan 1978, p. 289). This idea is supported by work on priming culture (context writ large) among bicultural participants. Incidental cultural icons can switch people from making causal attributions as members of a culture with "independent" self-concepts, to members with "interdependent" self-concepts (Markus & Kitayama 1991) and vice versa, quickly and unconsciously (Hong et al. 2000). Presumably cultural differences in moral judgment and responsibility attributions follow (e.g., Miller 1984 on differences between attributions by Indian and American children).

Further, Shimizu et al. (2017) showed that cultural differences in memory effects of "spontaneous trait inferences" reside entirely in automatic (unconscious, implicit) rather than controlled (conscious, explicit) processes. Spontaneous personality trait inferences are those made without intentions or awareness (Uleman et al. 2012). Because traits are commonly understood as causes (cf. Uleman 2015), this shows a process by which cultural (i.e., collaborative) differences in causal attributions may emerge. Using Jacoby's (1991) process dissociation procedure (PDP; see next paragraph), Shimizu et al. (2017) argue that most important cultural differences occur through automatic processes.

Dual process theories reflect many different dualities. In chapter 3 Doris highlights the one between automatic and controlled processing from semantic priming studies, in which these two processes yield conflicting results in cognition or behavior. Note that there are many kinds of priming (e.g., repetition, goal, procedural, and perceptual), each with its own properties (e.g., Förster et al. 2009), although this does not alter Doris's main point. There are also many definitions of "automatic" and "controlled." Processes may be called automatic if people are unaware of them and/or unaware of their effects; they are not intended; they have short reaction times (<500 ms); they require little or no cognitive capacity; concurrent cognitive tasks do not interfere with them; or they are uncontrollable (see Bargh 1994). Jacoby's (1991) PDP differs by estimating how much control exists when participants actually attempt to control cognition or behavior. We prefer Jacoby's definition because it does not label processes automatic if they have any (but rarely all) of the above characteristics, which do not always co-occur. The errors that occur in the false recognition paradigm (that has become standard for detecting spontaneous inferences) occur in spite of participants' efforts to control the unconscious intrusion of these inferences into task performance (Todorov & Uleman 2004). And they provide the bases for estimating the simultaneous operation of automatic and controlled processes in spontaneous inferences.

Doris also recognizes the distinction between holding people accountable for behavior prompted by events of which they are not aware (e.g., subliminal priming) and events that operate uncontrollably in spite of awareness and goals (e.g., addictions). These are both taken as challenges to a reflective position. Several varieties of self-control and its absence are examined in a fine collection of 27 chapters edited by Hassin et al. (2010), from the neural to the social level.

"Selves" recognizes that there are many (Ch. 8) and that their effects differ. Selves define the threats and values that determine one's morally responsible agency. Remarkably, even the implicit accessibility of various selves affects thought and behavior. Blaming the victim (whether of rape, non-sexual assault, or natural disaster) is commonly attributed to the "belief in a just world" (BJW): believing that people deserve what they get and get what they deserve, and that the world is morally predictable. Rather than relinquish this belief when they are faced with

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

apparently unfair victimization, people may deal with their distress by blaming the victim, thereby restoring a predictable and fair world in which only the deserving encounter harm. However, research has produced inconsistent results: sometimes victims elicit compassion and support. In past research on victim blame, BJW and self-relevance were often confounded. So we manipulated them independently in a series of six studies (Granot et al., under review). We found that the classic phenomenon of blaming the victim only occurs when both justice concerns and relevance to the self are activated.

In all of these studies participants read vignettes describing a victimization: a newspaper account of a hurricane victim, a fictional account of a young adult assaulted after a party, and genuine accounts of a sexual assault and an armed robbery on campus. Relevance to the self was manipulated by asking participants to assume a first-person or a third-person observer perspective; by displaying a photograph of the participant or a confederate on the participant's computer screen throughout the study; or by presenting "personal safety tips" to half of the participants in the case of assaults. While these manipulations of self-relevance did not ask participants to adopt "selves" that were unfamiliar to them, they activated the self in both explicit and subtle ways.

In each study, and across studies in a meta-analysis, blaming the victim only occurred when both self and justice concerns were high. Morally responsible agency was differently attributed in the same vignettes on the basis of relatively incidental changes in self concerns. These findings not only clarify the basis for inconsistencies in prior BJW findings, but they also illustrate the contextual malleability of moral judgments.

Should anyone doubt that personal relevance affects morality judgments relative to a more evenhanded god's-eye-view, Ham and Van den Bos (2008) showed how the two perspectives can differ from each other even within one participant. In two studies, participants read brief vignettes describing unjust events. Some vignettes were more relevant to the self than others, for example, "You and your colleague do the same work. You make(s) 1400 Euros a month and your colleague makes 4100 Euros a month" (Ham & Van den Bos 2008, p. 700). Spontaneous justice inferences were measured through response times (RTs) to justice-related words in a probe-recognition paradigm, i.e., were measured implicitly. Justice concerns arose most strongly (i.e., RTs were slowed most) to unjust vignettes involving the self (or to a relevant friend in study 2), compared to vignettes involving strangers. That is, injustice activated justice concerns only under high self-relevance. Explicit judgments were also obtained from these same participants, of how just events were in similar vignettes. Here self- (or friend-) relevance made no difference. Thus, participants were simultaneously of two minds. Implicitly (spontaneously), self-relevance was taken into account in activating justice concerns, but explicitly it was not. Not only are there multiple selves, but also the same self can have different effects for implicit and explicit judgments.

Identity (Ch. 8) affects how visual information is encoded and processed, in ways directly relevant to morally responsible agency. Doris posits the possibility that cultural identities might lead people to attend to different aspects of agency, but even attending to the same things does not ensure such information is similarly processed. Granot et al. (2014) asked participants to judge responsibility and blame from videotapes of altercations between two parties. In some cases, these were from dashboard cameras of police officers stopping motorists for traffic violations; in other cases, these were staged fights purportedly from security cameras. Participants' identification with one of the parties was either measured through self-reports (identification with police) or manipulated through assigning the parties to otherwise neutral in-groups and out-groups. These identifications, interacting with attention, affected judgments of blame and responsibility in a counterintuitive way, one that Doris would recognize as difficult to justify. Those who viewed the judgment target briefly were unaffected by identification – all blamed the target similarly. But those who studied the target more thoroughly were polarized in their judgments of blame and responsibility. If the target was a police officer with whom one identified, or a member of one's in-group, blame and responsibility were less; if the target was the out-group party, blame and responsibility were greater.

"Collaboration" again. Is the expectation of logical consistency in judging morality culturally relative? Perhaps social collaboration produces logically consistent systems of moral thought, at least in local linguistic communities. Many prominent Western philosophers (e.g., Kant; see Uleman 2010) have sought to develop logically consistent systems of moral thought. But Nisbett et al. (2001) suggest this may be a particularly Western concern. They note broad cultural variations in systems of thought, and contrast holistic (traditional Chinese) with analytic (traditional Greek) thought:

> We define holistic thought as involving an orientation to the context or field as a whole . . . an emphasis on change, a *recognition of contradiction* and of the need for *multiple perspectives*, and a search for the "Middle Way" between opposing propositions. We define analytic thought as involving detachment of the object from its context. . . . Inferences rest in part on the practice of decontextualizing structure from content, the *use of formal logic*, and *avoidance of contradiction.* (Nisbett et al. 2001, p. 293, italics added)

Historically, Chinese culture valued holistic thought whereas Greek culture valued analytic thought and its requirement of logical consistency. Nisbett et al. (2001) cite extensive evidence for the persistence of this cultural difference between Eastern and Western systems of thought, with Westerners more concerned with resolving than transcending contradictions.

This suggests that "collaboration" may resolve logical contradictions only when the cultural "system of thought" requires a resolution. Otherwise, multiple perspectives are embraced and not found wanting. It can also be argued that moral reasoning (and reasoning in general) is in the service of self-justification (Mercier & Sperber 2011). While this is collaborative in that it necessarily involves others, it is unlikely to produce a god's-eye-view of moral responsibility.

Manipulation, oppression, and the deep self

doi:10.1017/S0140525X17000826, e61

Manuel R. Vargas

Department of Philosophy, University of California San Diego, La Jolla, CA 92093.

mrvargas@ucsd.edu **http://vargasphilosophy.com**

Abstract: This essay considers various kinds of manipulation cases (local and global, dispositional and situational), and how Doris's Deep Self-style theory of responsibility fares in light of them. Agents acting with preferences adaptively formed under oppression are an especially interesting challenge for this sort of view, and the article considers what options may be available to Doris and others.

According to Deep Self theories, an agent is an apt candidate for blame when she acts in accord with values, self-governing policies, or particular higher-order desires (Frankfurt 1971; Watson 1975; Bratman 2007; Sripada 2015b). Such theories are appealing because they allow us to distinguish between wayward or "alien" impulses and actions that reflect the agent's "true," "deep," or "real" convictions. They also capture the idea that one reason for blaming wrongdoers is that the wrongdoing expresses something about the wrongdoer. However, Deep Self views face an important objection: they deliver counterintuitive verdicts about moral responsibility in manipulation cases (McKenna & Pereboom 2016). To see why, consider an agent whose deep self is manipulated, unknowingly "implanted" with values or desires

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

that replace her prior desires and values. Such agents can seem to be paradigmatically not responsible for actions derived from the manipulation. But on a Deep Self theory, the basis of responsibility just is whether the action reflects or expresses the values (or what have you) that the agent has. So it looks like the Deep Self theorist has to say that action that flows from the manipulated self is still responsible action. Manipulation cases – and, as I'll argue, some related but less science-fiction-y examples, including oppression and adaptive preferences – are a deep problem for the Deep Self approach.

Enter John Doris's (2015b) wonderful recent book, *Talking to Our Selves*. Doris offers an appealing upgrade to the traditional Deep Self account of responsible agency. He holds that responsible agency is present when an agent's actions are structured by values, or desires the agent accepts as determinative in practical planning. However, there is no requirement that the agent be aware of these values, or have self-consciously adopted them. What is required is only that the agents be willing to appeal to those desires or values in the justification of action plans. In keeping with Doris's emphasis on collaborativism (or the idea that optimal human rationality is socially embedded), he maintains that values can be discovered and even created in the social context of collaborative reasoning.

Can Doris's account overcome worries about manipulation? My suspicion is that the social dimension of Doris's account raises challenges that are particularly difficult to address within the confines of the Deep Self approach. Here, I consider several different kinds of manipulation (global and local, disposition- vs. situation-focused). A range of real-world cases suggest that it is difficult for his account to capture some familiar convictions about when oppression undermines culpability.

Manipulation cases are not a primary concern of Doris's book, but he does remark on a case of global manipulation, where an agent is subject to comprehensive and coercive value indoctrination. The model here is Patty Hearst's kidnapping and subsequent participation in the Symbionese Liberation Army in 1974. Doris argues, plausibly enough, that his account can handle such cases. Coercive indoctrination, he thinks, is likely to bring with it impairments in valuational capacities and value-expression, and even disruptions of personal identity (p. 31). Either Hearst's actions didn't express her values (if, for example, she was coerced or impaired), and thus she wasn't responsible, or her actions did express her values (without impairment), but then she was responsible for that reason.

What about cases of local value manipulation? Imagine a person who values being generous with comments on student research across all of the usual contexts, but who is unknowingly subjected to a manipulation where that generosity is deleted – but only in workshop contexts and not in office hours, labs, or conferences. Is the post-manipulated person responsible for her insensitivity in the workshop context? I'd wager that many would say no. Or consider a case of a monogamous lover who, because of manipulation, now values infidelity, but only in a narrow context. Or, consider a once-relaxed commuter who, post-manipulation, is made particularly prone to road rage, but only on rainy days. Implanted or manipulated values seem like the wrong kind of basis for responsibility.

Akin to the global case, perhaps Doris will lean on the thought that upstream manipulations produce downstream impairments to valuational capacities (2015b, p. 32). It is an empirical question whether we will ever have the ability to narrowly manipulate an agent's values in the way I've suggested. It is also unclear why local manipulation would necessarily bring with it impairments to an agent's valuational capacities or personal identity, and if so, why those impairments would always be operative in just those contexts where the manipulated values are in play. At any rate, there is no obvious conceptual barrier to the possibility of a local value manipulation without impairment. To the extent to which we find local manipulation cases to be instances of non-responsibility, then the Deep Self theory gets the wrong verdict.

The cases described above have all been instances of disposition-focused manipulation, where what is manipulated is the agent's "in the head" psychological dispositions. What about "out of the head" manipulation cases, or situation-focused manipulations? Suppose I know you are desperate to feed your family and I offer you demeaning and potentially illegal work that I know you are only willing to take out of desperation and lack of other alternatives. Suppose, too, that we both know that the work will likely shift your values in a direction that better comports with that work. Suppose further that I have some control over whether you have access to more palatable alternatives, and have conspired to ensure that you don't have access to those alternatives. Would you be fully responsible for the choice to take that job, and all that follows? Whatever the right answer is – and I suspect intuitions differ about choices under exploitatively engineered contexts – the fact that we can wonder about such cases seems puzzling, given a Deep Self view. Situation-focused manipulations look like they should be entirely irrelevant to moral responsibility, at least on a traditional Deep Self theory.

Oppressive social contexts may help bring out the stakes of the underlying puzzle. Let oppression be the property of unjust or immoral treatment, social relations, or distributions of opportunities, when it is produced by immoral or unjust social and political arrangements (Vargas, forthcoming-b). Although oppression is not always an impediment to responsibility, it is sometimes part of the explanation for why it doesn't always seem to be the fault of desperate people when they do desperate things.

At first pass, there are plenty of things Doris might say on behalf of the Deep Self approach in contexts of oppression. Consider someone who reluctantly takes to, say, low-level drug trafficking because in his part of town the non-criminal ways of earning money are difficult to secure, involve considerable burdens (e.g., traveling through hostile neighborhoods and/or relying on lengthy and uncertain commutes), or come at particularly high social costs (risking estrangement and vulnerability to violence for not participating in peer-group activities). In such a case the reluctant dealer would not be acting in accord with his values, and to that extent would not be as culpable as he would be were he acting from his values. So here the Deep Self theory delivers the right verdict. Fair enough.

The challenge of situational manipulations is deeper, however. One way situational manipulations work is by modifying dispositions. Take the case of adaptive preferences. Adaptive preferences are preferences that are formed in response to restricted options (Elster 1983). The particularly troubling cases of adaptive preferences are when the preferences are for things that are either counter to one's flourishing or otherwise not what one would prefer under more normatively optimal circumstances (Khader 2011). So, for example, in cases of domestic violence the abused partner may come to think of the abuse as merited or deserved. Or someone might come to think that because of her social identity (gender, race, social class, etc.), her labor deserves less compensation than it would were it done by someone with a different social identity.

Social orders inculcate norms that advantage some at the cost of others, and oppression plausibly relies on internalization for much of its efficacy. Value formation frequently occurs under conditions where people have an inadequate opportunity to deliberate upon and to choose morally palatable alternatives. If so, then the worry about adaptive preferences and the effects of oppression more generally on culpability are not readily addressed by simply consulting the offending agent's deep self. The worry is just that in the real world, deep selves are too often the products of processes that are themselves culpability-undermining. Either we need a compelling error theory for these intuitions, or we need to give up the idea that responsibility is grounded in the history-insensitive valuational structure of an agent.

Doris could address these challenges by stipulating a historical condition on moral responsibility, as others have done (Fischer & Ravizza 1998; Mele 2009). However, this would be at odds with Doris's explicit strategy in the book, which eschews appeals to

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

history in favor of appeals to an agent's occurrent psychological features (Doris 2015b, pp. 30–31). More importantly, he'd need some explanation of why such a requirement isn't an *ad hoc* departure from the basic explanatory strategy of the Deep Self theory. History might matter, but if the way it matters is antecedent to the presence or absence of the Deep Self, one might wonder whether the Deep Self is merely symptomatic of something else that actually grounds responsibility.

A different response could build on the collaborativist/socially responsive element that animates Doris's particular approach. Perhaps Doris could maintain that in some sense, for all agents, it is adaptive preferences all the way down. If so, then there is nothing special about the apparently awkward cases; like all cases of responsibility, it is a matter of whether the putatively culpable action reflects the agent's deep self.

That's a principled reply, if a costly one. The spouse who thinks she deserves abuse and puts herself and her kids at risk would, according to such an account, be fully culpable because she takes her meriting abuse to be determinative in practical reasoning. The victim of wage and employment discrimination who fails to protest his treatment because he has internalized racial and class prejudice and thinks he doesn't deserve a well-paying job is acting responsibly, says the theory, if he acts from internalized values. Perhaps it is an insight from philosophical theorizing that such agents enjoy no diminution of their responsibility. It would take a compelling story to overturn the widespread sense that oppression and adaptive preferences matter for responsibility.

It is unclear how to square our evident willingness to find some values and their formation as an inadequate basis for responsibility with the two chief features of Doris's account of responsibility, i.e., acknowledgment of the way situations shape dispositions and the idea that responsibility is grounded in a Deep Self. I'm somewhat more optimistic about accounts that ground responsibility in rational capacities (Vargas, forthcoming-b). On such accounts, if the agent is insufficiently capable of recognizing and suitably responding to moral considerations, then wrongful actions (grounded in preferences adaptively formed under oppression) aren't instances of responsible agency for reasons of rational impairment. Mitigation or diminutions of responsibility are explained in terms of constraints on the ability of agents to recognize and respond to moral considerations. For Deep Self theorists, however, to appeal to this sort of rational impairment is tantamount to abandoning the Deep Self approach.

Doris could appeal to his pluralism about responsible agency (pp. 12, 171–75), addressing such cases by appeal to resources that are not, as it were, "deep selfy." Such a strategy would suffer its own cost: if Doris has to appeal to non-Deep Self theories to shore up the Deep Self account, then it gets harder to insist that the Deep Self approach is a particularly helpful way to think about responsibility.

I'm not sure what the right answer is here, but I've no doubt Doris will find an insightful way forward.

To kill a bee: The aptness and moralistic heuristics of reactive attitudes

doi:10.1017/S0140525X17000838, e62

Hugo Viciana,[a] Antonio Gaitán,[b] and Fernando Aguiar[a]

[a]Instituto de Estudios Sociales Avanzados, CSIC, Plaza Campo Santo de los Mártires, 7, 14004, Cordoba, Spain; [b]Facultad de Humanidades, Universidad Carlos III de Madrid, Calle Madrid, 126, 28903 Getafe (Madrid) España.

hviciana@iesa.csic.es www.hugoviciana.net
agaitan@hum.uc3m.es faguiar@iesa.csic.es

Abstract: Although we are sensitive to the advantages of reactive attitudes as a starting point, we are concerned that confusion on the level of analysis can easily plague this type of account. We argue that what is needed here is a serious appraisal of the effects on the promotion of values of moralistic responses toward different types of agency.

Following Peter Strawson, John Doris claims that the 'right way' of thinking about agency should attend to those practices where we tend to ascribe moral responsibility. These practices are usually signaled by the presence of reactive attitudes. Reactive attitudes (e.g., gratitude, resentment, indignation, anger, guilt) are peculiar kinds of emotions whose expression we recognize as proper for some typical and paramount interpersonal relationships (Scanlon 2013). This line of thinking highlights two points, which are half descriptive and half normative. Reactive attitudes are so deeply embedded in our psychological evolved nature and social interactions that attempts to revise these attitudes must be seen as carrying the burden of the proof in a cost-benefit analysis. Furthermore, insofar as reactive attitudes are affects and emotions aimed to regulate our behavior within a set of interpersonal relationships, they offer a natural path to ground normativity on a factual basis, including a genuine feeling of 'to-be-doneness' attached to them (Mackie 1977).

We agree on the fecundity of thinking about moral responsibility through this lens, and we find *Talking to Our Selves* to be one of the most refreshing books on these issues that we have read in recent years. However, we want to comment on the way Doris models agency through reactive attitudes. For Doris, reactive attitudes are symptoms of morally responsible agency. In particular, the aptness of a reactive attitude toward X is a symptom of X's agency. Thus, if you can justifiably express a reactive attitude toward X, then X will surely be a morally responsible agent.

We think that some aspects of this modeling, although rich and suggestive, are importantly undertheorized. In what follows, we will briefly develop two points on the relationship between reactive attitudes and agency. First, reactive attitudes can sometimes be apt even if they are not tracking morally responsible agency. Aptness of reactive attitudes and hence characterizations of moral responsibility seem to fluctuate between different levels of what "apt" is. Second, in his characterization of agency, Doris sidetracks this slippery slope of the aptness condition with a reference to the values toward which our emotional attitudes react. As we will see below, this is equally problematic.

As we have just mentioned, sometimes reactive attitudes are apt even when we are not tracking agency by exemplifying them. In fact, it has been recently pointed out that some societies tend to discount intentionality for purposes of assigning moral responsibility (Barrett et al. 2016). Doris himself recognizes this possibility by mentioning cases of 'strict liability' (Doris 2015b, pp. 24, 154–55). In cases of strict liability (e.g., warfare atrocities, catastrophic slips), we get moral responsibility – and associated reactive attitudes – directed toward targets that lack core features of agency (e.g., intentionality, knowledge of the outcomes associated to the relevant behavior). Additionally, many times, the object of a reactive attitude is not straightforwardly related to issues of merit and lack thereof (Levy 2011). We can love someone even if we recognize that he or she does not deserve our affection. The same general point might apply for admiration, respect, shame, pride, and so on.

To further examine the possibility of having reactive attitudes directed toward clear instances of non-agency, let us be a bit of a gadfly and argue against Doris's proposal with his favorite example of the bee. Our anger toward the bee could be, in fact, an apt reaction in terms of the goal we want to facilitate in our relation with the bee; that is, not being harmed by the bee or being undisturbed by insects while on a picnic. Even if the bee is clearly something of an agential blob in terms of moral agency, the anger we direct *against such an organism* (and not against an action which has not happened yet and we try to avoid) can be seen as an efficient way of facilitating the above goal.

Our point is simply that in judging the aptness of the reactive attitudes there is easily confusion of a sort that has already been noticed (d'Arms & Jacobson 2000). Although we assume Doris is not unaware of that danger, we consider the following under

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

a different angle. When judging the appropriateness of reactive attitudes it is not obvious how to adopt a neutral or impartial perspective. Elements of reactive attitudes (e.g., guilt, resentment, admiration, contempt) that are part of the examined system of responsibility practices might supply functions not only of a different kind but also at different levels. This implies that we should be extremely clear about the level from which we are appraising the aptness of a reactive attitude in a given case.

In its most basic form, there are at least two different levels of appraisal for evaluating the aptness of reactive attitudes. From the *individual level*, reactive attitudes are a key element for regulating our social interactions and for signaling our status in partner-choice selection in both biological (Debove et al. 2015) and cultural evolution (Fessler & Holbrook 2013). As Robert Frank (1988) claimed 30 years ago, moral emotions supply to the individual a selective tool for maximizing her chances of being included in mutually cooperative enterprises. Thus, moralistic tendencies are plausibly psychologically evolved tools apt for uses in intra-group competition (Gavrilets 2012).

Normative discourse in philosophy on the appropriateness of reactive attitudes usually takes place at a different level, which is also a higher level than the individual or even subpersonal level just described. It is usually at a *group level of analysis* in which the function of the responsibility system is the proper target of discussion. From this perspective, reactive attitudes would be of paramount importance in normative and evaluative processes of negotiation aimed at a group's stability and cohesion (how inclusive that group is, is another question). It is easy to recognize this default view on scholarly work on reactive attitudes, for instance, in Strawson's well-known appeal to the value of reactive attitudes for "human life." Returning to the case of the bee, feeling anger could be "fitting" for the person who, by following such a pedestrian strategy, promotes their goal of being undisturbed while having lunch in the park. Less clear is whether feeling anger toward a bee will be adaptive from a certain group standpoint (e.g., right now we are rather short of bees for pollination . . .). Shall this make us doubt Doris's premise that the aptness of reactive attitudes is a symptom of morally responsible agency? The general point, again, is that both levels of analysis can deliver opposite verdicts about the aptness of a peculiar reaction or trait. When Doris (2015b) writes, for instance, "there's pretty good reason for you to be angry with me for what I did, if what I did is a function of my mean-spirited matterings" (p. 169), he conflates how mean-spiritness is an attribute at the individual level of analysis in a partner-choice framework (e.g., "you don't wanna be friends with me pal?"), whereas the justification of the practices of the responsibility system should take place at a higher, group level of analysis (e.g., "should we accept and encourage anger among conflictual relationships in our group?"). We would not like to be uncharitable to Doris's account. Our point is rather that it is not easy to not succumb to this slippery slope of aptness. "*Aptness for whom?*" should be a slogan here as well.

Let's now attend to our second concern with Doris's view of the relationship between reactive attitudes and agency. In Doris's theoretical framework it is the "reference to each man's values [that] explain why they deserve the attitudes I subject them to" (Doris 2015b, p. 37). We believe, however, that such an emphasis on the primacy of values of the morally responsible agent is infelicitous. That is, the case not only because explaining human behavior values can sometimes be epiphenomenal to other characteristics that are not really under the control or guidance of the agent, but also because, more generally, reactive attitudes might be inappropriate responses to those very same values that excite them in an agonistic fashion. In other words, reactive attitudes can be a functional response to competing values (it might be their individual level function to respond when confronted with specific values) without these reactions being adequate to limiting or canalizing those competing values in socially desirable ways.

Consider the case of reactive attitudes inside the family. For millennia, it was a deep-seated belief that it was justifiable to corporally punish children, even using extreme violence. Cultural prevalence and persistence of proverbs in very different societies are a sign of the enduring attractiveness of this component of the responsibility system. "He that spareth the rod hateth his son. But he that loveth him chasteneth him betimes." "Better to beat your child when small than to see him hanged when grown" (see Pinker 2011). The moralistic rationale of these kinds of practices was to protect a hierarchical relationship between children and other adults. Such a relational model was presumed to be essential in forming the underage being (Fiske & Rai 2014). To state it delicately, no meta-analysis of the available scientific evidence has found any positive effect of this kind of moralistic violence either on children's development or on the quality of the relationship between children and their elders.

Distinguishing among non-converging levels is a crucial step in understanding that social and cultural evolution might cause certain reactive attitudes to appear as justifiably more apt without that being a real proof of their efficacy against the values that trigger them. For instance, Jennifer Jacquet (2015) underlined how recent social trends have promoted the reactive attitude of guilt when facing environmentally pernicious consumerism. Jacquet argued that such emotional sign of the times is not very efficient in solving large-scale cooperation dilemmas. Instead, coordinated practices of shaming the most environmentally pernicious agents would prove much more effective. Laboratory experiments in cooperative dilemmas suggest that contribution to public goods can be heightened up to 50% when the practice of shaming uncooperative agents remains a possibility (Jacquet et al. 2011). Private and public initiatives conducive to shaming the most disruptive agents in real-world settings have actually led to impressive results in saving water or decreasing tax fraud.

The *"apt for whom?"* question can also help in visualizing what could be termed the Machiavellian challenge in the appropriateness of moral attitudes when reacting against other agents' values. Indeed, there is a natural tendency toward the formation of coalitions (Weeden & Kurzban 2014) that claim that certain reactive attitudes against other groups are justified while defending their inclusive interests. Even though those coalitions can be broad and end up promoting something akin to the "public interest," a theory of morally responsible agency must be reflective on these issues. The impact of the widespread justification of responsibility practices on public policies is well-documented. Reactive attitudes toward merit, effort, and luck are strong predictors of different levels of redistributive social spending across a wide range of societies (Alesina & Angeletos 2005; Benabou & Tirole 2006). To sum up, reactive attitudes, because they are biologically and evolutionarily anchored on an individual (and even subpersonal) rationality, do not necessarily react efficiently against other values that excite them.

When drawing conclusions from these considerations one could remind oneself not only that values evolve through economic and technological pressures (Morris 2015) – and here we applaud Doris when he points out that "[morally responsible] agents are negotiations" (2015b, p. 148) – but also it is essential to keep in mind that some coalitions and groups can impose a disproportionate burden of the responsibility system on other groups. A good characterization of moral agency should make room for this fact. Think simply of how motherhood, as compared to fatherhood, plausibly has been and continues to be accompanied by very demanding responsibilities and expectations. Or think simply about how legal and judicial systems have been periodically criticized since the Enlightenment for imposing stronger responsibility requirements on some groups than others. As we can grasp how these demands and requirements have evolved, shall we conceive of morally responsible agency as evolving as well?

ACKNOWLEDGMENT

Funding for this work comes from the project FFI2015-67569-C2-2-P, MINECO/FEDER: "The constitution of the subject in social interaction. Identity, norms and the sense of action from an empirical perspective."

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

"Defeaters" don't matter

doi:10.1017/S0140525X1700084X, e63

Zina B. Ward[a] and Edouard Machery[b]

[a]Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA 15260; [b]Department of History and Philosophy of Science and the Center for Philosophy of Science, University of Pittsburgh, Pittsburgh, PA 15260.

zina.b.ward@pitt.edu machery@pitt.edu
http://zinabward.com www.edouardmachery.com

Abstract: We argue that the exercise of agency is compatible with the presence of what Doris calls "defeaters." In order to undermine reflectivist theories of agency and support his valuational alternative, Doris must not simply show that defeaters exist but rather establish that some agentive behaviors *do* express a person's values *without* involving reflection.

Doris's skeptical challenge and proposed account of responsible agency in his provocative book, *Talking to Our Selves*, are built around what he calls "defeaters": "causes of [an actor's] cognition or behavior [that] would not be recognized by the actor as reasons for that cognition or behavior, were she aware of these causes at the time of performance" (Doris 2015b, p. 64). We agree that human life is shot through with defeaters, so understood. Especially if causes count as defeaters even when they only have a small impact on behavior (and Doris argues they do), the diversity and ubiquity of defeaters is even more staggering than Doris himself acknowledges. They include not just the causes of curiosities like the Watching Eyes Effect, implicit egotism, and the Ballot Order Effect, but also a whole host of more mundane factors – good news for Doris because at least some of these curiosities are likely to be false positives (e.g., Northover et al. 2017 on the Watching Eyes Effect). People's cognition and behavior is influenced by their cultural background (Nisbett et al. 2001), upbringing (Cherlin et al. 1991), personality type (Back et al. 2009), mood state (Gardner 1985), hunger (Bushman et al. 2014), etc. In some cases, the actor would recognize these causes as reasons for her behavior; in many, however, she would not, and they would count as defeaters.

For instance, consider Reyna, whose decision to use reusable grocery bags at the store is causally influenced by the fact that she grew up in Portland, Oregon; that she is highly extroverted; and that she was raised by a hippie. (None of these factors makes all the difference, we may suppose, but they all contribute in the sense that hometown, personality type, and parental attitudes have a small but significant effect on reusable bag use.) Nevertheless, Reyna probably wouldn't say that being from Oregon, having an extroverted personality, or being the child of a hippie is a *reason* to use reusable bags. As Doris (2015b) puts it, these are "rationally and ethically arbitrary influences" on her behavior (p. 64).

But Doris is not a skeptic about human agency; indeed, he wants to save it in the face of apparent defeaters. A central task for a theory of agency, he claims, is to rule out the presence of defeaters in putative cases of agentive behavior (Doris 2015b, p. 66). His valuational account of agency is supposed to accomplish this: "[W]hen we are justified in asserting that a person's conduct expresses her values, we are justified in ruling out the presence of defeaters, and are therefore justified in attributing an exercise of agency" (p. 159).

But this can't be right. Even when an action expresses the agent's values, it must still be influenced by a myriad of causes that she would not recognize as reasons, were she aware of them. For instance, Reyna's use of reusable bags expresses her liberal values, which are *themselves* influenced by the place of her upbringing, among many other rationally arbitrary causes. Doris is sensitive to the possible mediation of such causes by reflection and seems to restrict the notion of a defeater to unmediated influences. As he puts it early on in *Talking to Our Selves*, "Best to distinguish cases where happenstance engages rational capacities from cases where happenstance bypasses rational capacities. I've been worrying about cases of bypassing: influences that are not vetted by rational capacities." And then: "It is these [bypassings] that must be ruled out . . . We do well to acknowledge that the origins of behavior are complex, and will often include any number of mediating factors" (Doris 2015b, pp. 72–73). These remarks should be extended to value expression as well. The category of defeaters must not include rationally arbitrary influences that are mediated by an agent's values. But this restriction isn't enough: Even behavior that expresses individuals' values is likely to be influenced by rationally arbitrary causes that are not mediated by them. Even if Reyna is expressing her value for environmental stewardship, that doesn't mean her behavior isn't *also* influenced by her extroverted personality. Or, to use Doris's own example, you may vote for the first candidate on the ballot because it is an expression of your values *and* because you are susceptible to the Ballot Order Effect. ("Because" here just implies that the factor has a nonzero effect size with respect to the behavior in question.) Surely Doris does not deny that behavior, like pretty much everything else, is multiply determined.

In short, value expression is entirely compatible with the presence of so-called defeaters. The valuational account of agency should therefore only require that, in order for an agent to be morally responsible for an action, the action must express *at least one* of the agent's values. It may also reflect numerous arbitrary causal influences (mediated or unmediated) that the agent would not recognize as justificatory; but that is not enough to disqualify it as agentive.

This friendly amendment to Doris's valuational account of agency raises the question: Aren't the same moves available to the reflectivist? Doris characterizes reflectivism as the doctrine that "the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do" (Doris 2015b, p. 19). Its corollary, he claims, is the idea that "the exercise of human agency requires accurate reflection" (p. 19). Before discussing reflectivism further, let us note an ambiguity in Doris's characterization: The idea that human agency rests in some way upon our capacity for reflection is distinct from the claim that every individual instance of agentive behavior must be preceded by reflection. Doris takes the latter as his target, but we suspect that at least some of the authors he cites are only committed to the former. For the purpose of this Commentary, however, we'll go along with Doris in taking reflectivism to be the claim that reflection precedes every exercise of responsible agency.

It would seem that the reflectivist can adopt clarifications analogous to those that Doris's valuational account of agency requires. First, she can emphasize that many putative defeaters are in fact mediated by reflection, and so don't count as genuine defeaters. If Reyna's Oregonian childhood prompts her to reflect on the environmental consequences of her actions and thus to use reusable bags, her hometown would not be a true defeater. As noted above, Doris gestures at this response but seems to think it does not apply to his cases (Doris 2015b, pp. 71–73). We suspect, however, that some of the empirical findings that he cites result from mediation of this kind. For instance, the Ballot Order Effect may be caused by voters reflecting for a diminishing amount of time on the accomplishments of each candidate as they scan down the ballot (Hogarth & Einhorn 1992; Krosnick et al. 2004). If the effect arises because the order of names influences the amount of cognitive resources that voters devote to reflecting on each one, the result does not undermine reflectivism as Doris intends. Second, the reflectivist can argue that the exercise of agency requires only *some* "self-conscious reflection about what to think and do," and that it is compatible with

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

the presence of other rationally arbitrary influences on behavior (Doris 2015b, p. 19). These two modifications exactly parallel the modifications to the valuational theory of agency proposed above.

We have little interest in reviving reflectivism, at least as Doris understands it. Rather, the point is this: presence or absence of rationally arbitrary influences on a behavior is orthogonal to the question of whether it constitutes an exercise of agency. Rationally arbitrary causes can be present simultaneously with reflection or with value expression. Thus, "defeaters" is in fact not a very appropriate name: Rationally arbitrary causes by themselves do not defeat responsible agency.

More importantly, once we recognize this point, it becomes clear that empirical research on defeaters per se is not what Doris needs to undermine reflectivism and support his valuational account of agency. Instead, Doris must show that some behaviors that are genuine exercises of agency *do* express a person's values *without* involving reflection. Let's call this task "the valuationist's challenge." Doris has too little to say about the positive part of the challenge. After raising doubts about whether a diverse array of behaviors are agentive, he does not go on to show how those behaviors are expressive of people's values, and thus safe from the specter of skepticism. The negative part of the valuationist's challenge is to show that some presumed exercises of agency are brought about with no reflection at all. It is not sufficient to argue (as Doris convincingly does) that agents often fail to reflect on *some* of the causes of their behavior.

Meeting the valuationist's challenge, particularly its negative part, is likely to be difficult because the reflectivist can appeal to a conception of reflection that falls short of undertaking an episode of conscious practical reasoning in which various alternative courses of action are considered and one is deliberately selected. True, if *that* is the type of reflection reflectivists are committed to, then some genuinely agentive behaviors take place without reflection, but that battle isn't worth fighting: A few minutes of introspection reveal the shortcomings of reflectivism so understood. Perhaps reflection involves instead "bringing to mind ideas or images meant to have some rational relation to the topic being considered, in the service of reaching a conclusion about what to think or do" (Arpaly & Schroeder 2012). Or perhaps it only requires paying attention to task-relevant cues and acting on that basis (Wu 2014, Ch. 3). On these understandings of reflection, Doris's task would be to provide empirical evidence that no such ideas or images are brought to mind, or no attention is paid to any relevant cues, prior to some apparently agentive behavior. The vast social-psychology literature on rationally arbitrary influences on behavior is silent on these questions.

Even more minimally, a dominant tradition in the philosophy of action is based on the idea that the exercise of agency requires that one act for the sake of reasons (Anscombe 1957; Raz 1999). Recently several authors have argued that non-conscious processes may be "responsive to reasons *qua* reasons" and thus give rise to intentional or rational action even in the absence of deliberation (Railton 2009; see also Railton 2006; Arpaly & Schroeder 2012). To dispute this view, Doris would need to demonstrate a total absence of reason-responsiveness in some behavior we would want to call agentive. While Doris may respond, perhaps reasonably, that this minimalist account of agency is not reflectivist, and therefore not his target, this response would be less convincing for the views mentioned above.

The fulfillment of the valuationist's challenge requires a clearer understanding of what reflection is and the presentation of apparent exercises of agency without it. Demonstrating the presence of "defeaters" is not enough, because the ubiquity of non-mediated causal influences on behavior ought to be acknowledged by proponents of both reflectivist and valuational accounts of agency.

The practice of everyday life provides supporters and inviters of morally responsible agency

doi:10.1017/S0140525X17000851, e64

Jörg Zinken[a] and Vasudevi Reddy[b]

[a]*Institut für Deutsche Sprache (IDS), 68161 Mannheim, Germany;*
[b]*Department of Psychology, University of Portsmouth, Portsmouth, United Kingdom PO1 2DY.*
zinken@ids-mannheim.de vasu.reddy@port.ac.uk
http://www1.ids-mannheim.de/prag/personal/zinken.html?L=1
http://www.port.ac.uk/department-of-psychology/research/situated-action-and-communication/

Abstract: Drawing on research from conversation analysis and developmental psychology, we point to the existence of "supporters" of morally responsible agency in everyday interaction: causes of our behavior that we are often unaware of, but that would make good-enough reasons for our actions, were we made aware of them.

Doris is troubled by "defeaters": causes of our behavior that we are not aware of and, to make things worse, that we would not think of as good *reasons* for our behavior were we made aware of them. Research in experimental social psychology claims to have identified many such defeaters. To cite one of Doris's examples, some people seem to vote for the guy at the top of the ballot, although they would probably not claim that "because he was top of the ballot" was the reason for their choice, or that it would make a *good* reason for anyone's choice. But if our decisions can be swayed in such irrational ways, the worry follows, how can we think of ourselves as responsible agents?

To soften the blow dealt to our agency by such potential "defeaters" and to support the overall thrust of Doris's work, we want to draw attention to research on everyday social life. Some of that research points to the existence of what we might call "supporters": causes of our behavior that we are often unaware of, but that would make good-enough reasons for our actions, were we made aware of them. Work in Conversation Analysis, which examines the normative order of our mundane social interactions, has identified many such supporters (e.g., Stivers et al. 2011; Zinken 2016).

Consider the ways a person might address a request for some action to another. Requests for small-scale acts of cooperation are ubiquitous in everyday life, and they provide a central arena for morally accountable agency. In languages around the world, we find complex systems of practices for making requests (Floyd et al. 2014). But how do I come to the decision to use an interrogative in one situation (*Can you pass me a plate?*) and an imperative in another (*Pass me a plate*)? As it turns out, we make these little decisions in highly systematic ways: Imperative requests – the predominant format for requesting little acts of cooperation across languages – build on the other's availability for the job, whereas interrogative requests take into account the fact that the other is occupied with something unrelated. In other words, a cause for choosing an interrogative over an imperative request is the relative lack of *continuity* of the requested action with what the other is already doing (e.g., Rossi 2012; Wootton 1997).

People seem to care about these contextual cues a lot. If your partner is standing in just the right spot to pass you the plate you want, you might begin to address him with an imperative. But if he begins to walk to the fridge right as you start talking, you might break off and restart your request in an interrogative format (data discussed in Zinken 2016). But when we asked videotaped participants about their choices in such everyday interactions, they either spoke broadly about matters of politeness, or suggested that their choice of request form was quite arbitrary. It seems then, and maybe unsurprisingly so, that we are not always aware of the subtle contextual cues that sway our decisions in our everyday interactions with others. But if we were made aware of them, would there be cause for embarrassment? I would say not: My more or less subliminal awareness of the fact that my partner's moving to

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

the fridge constitutes an engagement that is in conflict with passing me a plate has moved me to express the value of respecting another person's autonomy. Two seconds earlier, my sensitivity to my partner's availability for cooperation had moved me to express another important value: our need to express closeness to others (Brown & Levinson 1987). In sum, the contextual cues that regularly sway our decisions in designing interactional moves can indeed make good reasons for occasions when we are asked to rationalize what we did (see Doris 2015b, pp. 141–43). They are supporters rather than defeaters of our morally responsible agency.

Surely we do have our irrational moments. But maybe the extant research in social psychology over-advertises these a bit. Research on everyday adult interaction instead finds "supporters" of moral agency to be ubiquitous in our social lives. And from the earliest months of infancy, we are drawn into action by "inviters" of action and response, without which development would be impossibly difficult. Infants begin from 6 or 7 months of age to understand and comply with adult directives. However, rather than being enabled by a newly developed grasp of others' intentions, infants come to this gradually. The adults around set up contexts of repeated and routinized invitations to act, to attend, to show off their skills to visitors, and to join in easy cultural rituals (Reddy et al. 2013). Infants are drawn into participating in these engagements, and without direct focus on it, begin to grasp the very structures of moral participation in social life.

These inviters and supporters seem to fit well into the theory of the social genesis and development of morally responsible agency that Doris has argued for. Their recognition might further reduce the appeal of skepticism about responsibility and agency.

Author's Response

Collaborating agents: Values, sociality, and moral responsibility

doi:10.1017/S0140525X17001935, e65

John M. Doris

Philosophy-Neuroscience-Psychology Program, Philosophy Department, Washington University in St. Louis, St. Louis, MO 63130

jdoris@wustl.edu **http://www.moralpsychology.net/jdoris/**

Abstract: I respond to the *Behavioral and Brain Sciences* commentaries on my book, *Talking to Our Selves: Reflection, Ignorance, and Agency*. I defend and amend both the skeptical challenge to morally responsible agency, that is, the book's impetus, and the anti-skeptical theory I develop to address that challenge. Regarding the skeptical challenge, I argue that it must be taken more seriously than some of my sanguine commentators assert, and consider some ways its impact might be blunted, such as by appeal to individual differences and the practical efficacy of human behavior. Regarding my positive theory, I defend the role of values in morally responsible agency against numerous criticisms, and consider various suggestions for elaborating my social, "collaborativist" account of morally responsible agency. In closing, I comment on the appropriate aspirations for theorizing about moral responsibility and agency.

Talking to Our Selves (Doris 2015b) attempts to make sense of some puzzling phenomena: cases where people's rational capacities are "bypassed" (Nahmias 2011, pp. 560–63), and the most perspicuous psychological explanations of their behavior do not include – or include enough, or centrally enough – considerations they would take as justifying reasons for their behavior. I understand such phenomena as intimating skepticism about morally responsible agency ("agency" for short),[1] a skepticism I develop in Part I of the book, and attempt to forestall in Part II.[2]

My responses to the commentaries collected here can usefully be organized in a similar way. In the first section I reconsider the skeptical challenge, together with some reactions to it that did not figure prominently in the book. In the second section I reconsider my anti-skeptical theory, with particular regard to its "valuational" and "collaborativist" aspects (collaborativism is probably the part of my approach with which commentators are most sympathetic). Finally, I close, in the third section, with brief remarks about the appropriate aspirations for theorizing about agency and responsibility. This organization does a tolerable job of imposing order on an impressively diverse and perceptive set of commentaries; unfortunately, it does not allow me to transcend limitations of space, and remark on every issue deserving of attention.

R1. The skeptical challenge

A number of commentators contend that I have oversold the skeptical challenge: some think the problem is readily solved, and some even doubt there's a problem that needs solving. On the other hand, numerous distinguished contributors to the agency literatures agree that the challenge is one theories of moral responsibility and agency should address (e. g., Arpaly, forthcoming; Fischer, forthcoming; Kane, forthcoming; Nelkin, forthcoming; Shoemaker 2015; Tiberius, forthcoming; Vargas, forthcoming-a) – an assessment shared by numerous of the present commentators, such as **Dunning**, who says there are "stark" contradictions between philosophical accounts of moral responsibility and the findings of contemporary psychology and cognitive science (cf. **Alfano**; **Beal & Rochat**; **Maibom**; **Mercier**; **Niemi & Graham**; **Pe-Curto, Deonna, & Sander** [**Pe-Curto et al.**]; **Sommers**; **Taylor**). Obviously, I side with the concerned, but it is worth considering why some interpreters are more sanguine.

R1.1. Skepticism and surprising effects

The skepticism I develop is empirically motivated, and is substantially dependent on the credibility of the science from which it draws. In both the book (pp. 44–49) and the précis, I discuss the recent replicability controversy in psychology, and I won't reprise my remarks here. But **Lambert & Dennett** mention a selection effect in the production of science that should be addressed. Psychologists don't get paid, to borrow their example, for demonstrating that people prefer $1,000 to a pin prick: "generating small, surprising effects may be rewarded in psychology," they say, "but it is not clear whether or how the common lore of everyday psychology that psychologists never bother to investigate is undermined" by this enterprise. Reward structures in academic psychology reflect, to some degree, the Surprising Effect Bias (p. 47), and like Lambert & Dennett, I'm confident that where psychologists set out to demonstrate Obvious Effects, they'll have decent luck in finding people behaving expectedly, rationally, and, perhaps, agentially.

Does this mean the surprising effects I recount are the unlikely exceptions that prove the agential rule? Maybe

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

not. Unexpected effects appear regularly throughout psychology (Roediger & Butler 2011), which makes me think that they are comparatively easy to get, and are not merely an artifact of the Surprising Effect Bias. I'm therefore guessing we should expect unexpected phenomena in life, where stimuli are likely to be vastly more various and potent than in controlled laboratory experiments.

While it is repeatedly observed that effect sizes in psychology tend to be of a size conventionally regarded as small (Cohen 1988, pp. 77–81), it should also be noted that the manipulations themselves are often rather slight, with nothing like the potency of real-world stimuli. The pictures used for inducing disgust in the lab, for example (e.g., Cameron et al. 2013), are the palest simulacra of something *truly* disgusting, like a dumpster festering with rot and squiggling with maggots on a sweltering August day. So, too, the "prison" in the Stanford Prison Study is hardly more than a parody of an actual correctional facility.

Given the limited force of many laboratory manipulations, isn't the small size of many unexpected effects exactly what we should expect? The remarkable thing, one might say, is not that the surprising effects are small, but that they can be gotten at all. By the same token, shouldn't we expect larger impacts from the rough and tumble of real-world influences? All this disinclines me to conclude that bypassing defeaters are largely an artifact of the Surprising Effects Bias.

R1.2. Skepticism, effect size, and defeaters

Fowers, Anderson, & Lang (**Fowers et al.**) contend I have overstated the skeptical challenge. I suspect this conviction can be traced to their understanding of the philosophical enterprise: "philosophers incline to search for absolute truth, which fuels the skeptic's challenge." Certainly, some philosophers, among them many historical Greats, incline to absolutism, but this is emphatically not my orientation; as I said in the book, I'm making a "wager" in conditions of scientific uncertainty, not searching for absolute truth (pp. 12–14, 48–49; cf. Machery & Doris, forthcoming). Moreover, my epistemic orientation is explicitly *fallibilist* (p. 65), meaning I think knowledge has its basis in *defeasible* justification, not "absolute" justification. The skeptical challenge is not ginned-up by philosophical absolutism.

Fowers et al. also misunderstand how the skeptical argument works. They say the experimental effects are "generally mild and inconsistent," and we are in at least partial agreement here; I acknowledge (pp. 61–64) that the relevant effects will often be small (while observing that small effects can be practically and/or theoretically important). But Fowers et al. also say "there is a small probability that a given individual act will involve 'bypassing' the actor's reason," and conclude that this "small" probability (whatever it is) means that the skeptical worry does not go through;[3] rather, the most we can conclude is that "humans are imperfect moral reasoners."

The skeptical argument does not require that the probability of defeaters be greater-than-small, but rather that the skeptical hypothesis be a *live* one, that requires ruling out (p. 66). If **Fowers et al.** could show that the probability of a defeater obtaining was *trivially* small, like the odds of winning a multistate lottery (lottery paradoxes noted!) maybe the skeptical argument would be a merely academic exercise. But that isn't what they show, nor does it seem to be entailed by their guesstimate that the probability of defeaters is small.

When **Fowers et al.** turn to my suggestion that defeaters may aggregate to undermine agency (Doris 2015b, p. 64), they say "it is just as likely that situational factors cancel one another out as cumulate" in a way that undermines agency. I'm not sure what supports their estimate. But suppose they are right, and it is exactly as likely that an aggregate canceling of defeaters obtains as it is that an aggregate defeater obtains. On this supposition, it is equally likely that agency is, or is not, undermined, and attributions of agency are the epistemic equivalent of flipping a coin; here, Fowers et al.'s scenario just is a skeptical scenario.

May presses the issue of effect size with an elegantly posed dilemma: "influences on many choices tend to be either substantial or arbitrary but not commonly *both*." For something to threaten agency, on May's picture, it would have to be simultaneously rationally arbitrary and a substantial influence on behavior, but most of the effects I mention, he thinks, are either insubstantial or non-arbitrary. Without the other, neither condition unsettles agency, so there aren't enough troubling effects to make trouble.

My argument focused on arbitrary influences (pp. 54, 64); roughly, influences that make unlikely justificatory material, like Ballot Order Effects. But, **May** argues, many of the experimental effects, like certain bystander effects, "aren't necessarily" considerations "we'd reject as non-reasons." True enough; being in a hurry, for example, might reasonably, in some contexts, count as a reason. But that a consideration can serve as a reason in some contexts does not mean it can't count as a defeater in others. If being in a hurry leaves you unresponsive where you otherwise would have helped, given your values, it's plausible that something has gone wrong for your agency, even if being in a hurry might have served you as a justification in another context, or for a person with different values. So, rationally non-arbitrary considerations sometimes count as defeaters.

It's not clear how to gloss "substantial," but *makes a practical difference* is a plausible reading. Again, the effects I'm considering often are often small, but even small effects of the sort typical in much psychology can make a practically meaningful difference, as they do in medicine (pp. 61–64). I suspect that most complex psychological effects are the function of many individually small influences, an observation I dub the *lotta-little principle* (Doris, in preparation). The truth of this principle does not give us reason to think that none of these effects can make a practical difference. No straw by itself is *the* straw that breaks the camel's back. But each straw, given all of the other straws, has an effect.

Once more, I'm not sure how one could substantiate **May**'s assertion that the combination of arbitrary and substantial doesn't "commonly" occur. He's surely right that "many of our choices are influenced by good reasons," but this is compatible with many choices being undermined by defeaters: that there are many short people is compatible with there also being many tall ones. Furthermore, it's not obvious to me how much confidence "not commonly" should yield, even if that estimate be granted. Everywhere I've lived in the United States, you see red-shouldered hawks a lot less than you see red-tailed hawks; the former, tragically, may be in decline, while the latter are still ubiquitous. That is, red-tails are "common;" red-

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

shoulders, not so much. Still, red-shoulders are common enough to reduce confidence in identification when one sees a buteo at sufficient distance in places where both birds range. Even an uncommon occurrence can have epistemic import.

Because I'm reticent about speculating frequencies, **May** wonders if "we didn't need all of the empirical evidence," and might just as well have generated skepticism by "imagination alone," as is done in the many philosophical intuition pumps featuring such unlikely scenarios as Epistemically Malicious Demons or Envatted Brains. Well, I use the empirical evidence as a kind of "possibility pump," elevating defeaters from a mere possibility that might be reasonably discounted, to a live possibility that requires ruling out. Interestingly, some skeptical intuition pumps famous in epistemology may likewise draw their force from empirical observations. For example, early on in his *First Meditation* (Descartes, 1641/2008, paragraphs 3–5), Descartes mentions several possibilities that, while perhaps not commonplace, happen often enough to give pause: perception is inaccurate or distorted, people hallucinate, people are unsure whether they're dreaming. I speculate that if such empirical possibilities were unknown, and there were no evidence of inaccurate perception, the imaginings induced by the skeptical intuition pumps would lose much of their power to provoke. Matters are the same, I think, for the contemplated agency skepticism; if the curious disruptions of agency I discuss were empirically unheard of, the skepticism would have a weaker bite.

Levy is likewise skeptical about the skeptical problem, which he dismisses as "largely illusory."[4] He argues that questions of agency are decided not by the reasons we *do* endorse, or *would* endorse were we aware of them, but rather by those "we *should* endorse." What we should endorse, according to Levy, is often fixed by what was "adaptive in our ancestral environment," because "a large proportion of these processes continue to track reasons." Therefore, he thinks, "very many, perhaps the overwhelming majority, of the processes that Doris identifies as defeaters are better seen as realizers of our agency than as defeaters of it."

I do not deny that there are many cases where automatic processes support rather than subvert agency (p. 51); some goals are better accomplished by automaticity. But many does not get us to "overwhelming majority," and yet again, it's quite unclear how the frequency of defeaters in the great corpus of human behavior can be confidently estimated.

There's more trouble for **Levy**'s suggestion than that. Adaptation, infamously, is not overly troubled by the demands of morality and rationality: do selfish genes make reasonable persons, who generously donate to worthwhile charities and prudently save for a secure retirement? Take the Cinderella Effect – stepchildren are more likely to be mistreated than are biological children, and this circumstance may well be an adaptive product of natural selection (Daly & Wilson 1996; 2007). But surely the existence of this "adaptive" tendency does not generate a moral prescription to mistreat stepchildren. Such examples are among the many reasons that attempts to find an evolutionary basis for ethics are often met with well-considered suspicion (e.g., Machery & Mallon 2010).

Levy's suggestion is problematic not only with respect to morality, but also with respect to agency. I formulated my account in terms of "subjective" reasons, rather than "objective" reasons, and maintained that the question for agency concerns whether the agent *herself* would treat the causes of her behavior as justifying reasons, not whether these reasons are reasons she ought have from some perspective other than her own (pp. 43–44, 70, 135). If a parent is committed to treating all of their children equally, but is biased against their stepchildren as a result of the Cinderella Effect, there are questions about the extent to which they're exercising agency. Their behavior runs counter to *their* values, and is not regulated by *their* reasons. The fact that there is some perspective which counts them as acting on good reasons, however abhorrent those considerations are to them, does not make it the case that they are acting agentially. Of course, one might seek to build an account of agency around a theory of objective reasons, the apparent agential significance of the actor's subjective perspective notwithstanding. But even if one takes on such demanding work, they should, by dint of familiar examples like the Cinderella Effect, hesitate to source reasons in adaptations.

Hirstein & Sifferd also propose estimated frequencies as a bulwark against skepticism: inaccurate self-awareness, they allow, is "fairly common," but inaccuracies that go uncorrected because the subject *cannot* correct them are less common. They insist that only the latter, less common cases, make trouble for responsibility: "in cases where we have the capacity to correct for our mistaken perceptions, using our brain's prefrontal executive processes, it would seem we are responsible for them." Because they think the responsibility-threatening cases where people lack this capacity to correct are pretty rare, Hirstein & Sifferd think the skeptical challenge is less serious than I suggested.

I won't say more about anti-skeptical frequency guesstimates, but I will say something about **Hirstein & Sifferd**'s appeal to capacities, because I argued in the book that this expedient cannot disarm skepticism (pp. 37–40). My argument is pretty simple: in many cases, folks occupy excusing conditions, and are therefore not responsible, when there is little doubt they have the responsibility-relevant capacities. Certainly I have the *capacity* to act another way when manipulation or coercion undermines my exercise of agency, it's just that circumstances prevent me from *exercising* that capacity. Indeed, here lies a way of understanding the distinction between excusing and exempting conditions (Doris 2002, pp. 129–30). Excuses are at issue when the actor is assumed to have the capacities requisite for agency, but are somehow prevented from exercising them, while actors in exempting conditions are supposed to lack the relevant capacities altogether, and are globally non-responsible. Thus, the presence of responsibility-relevant capacities is not sufficient for the attribution of responsibility, because people in excusing conditions, who are paradigmatically (though locally) non-responsible may be assumed to have them.

A promising approach, which I should have considered in the book, links responsibility to fair opportunity (Brink & Nelkin 2013). On this picture, when I'm in excusing conditions I have the capacity to exercise agency, but factors like coercion and manipulation deny me fair opportunity to do so. Here, if defeaters are to count as underminers of responsibility, it must be the case that their subjects lacked fair opportunity to resist them. And here, self-ignorance can matter, because it is not clear that people have

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

fair opportunity to resist influences of which they are unaware. One might disagree about this – if your vote is really a vote of conviction, maybe you can fairly be expected to resist ballot order effects, whether or not you know about them. But working this out will involve more than simply observing that the subjects of defeaters have the capacity to resist their influence.

Unlike other anti-skeptical commentators, **Ward & Machery**'s resistance is not based on guesstimates of defeaters' rarity. Indeed, they allow the possibility that "the diversity and ubiquity of defeaters is even more staggering than Doris himself acknowledges," and recognize, as I do (p. 72), that the causal history of any behavior is likely to be shot through with influences those so influenced would be unlikely to treat as justifying reasons. What Ward & Machery doubt is that these influences threaten agency: there are huge numbers of what I term defeaters, but according to them, most of these "so-called defeaters" aren't agency-undermining. As they observe, "even behavior that expresses individuals' values is likely to be influenced by rationally arbitrary causes that are not mediated by" their values; if these causes have to be counted as defeaters, there would seem to be an objectionable scarcity of agency.

Ward & Machery propose this fix: the "valuational account of agency should . . . only require that, in order for an agent to be morally responsible for an action, the action must express *at least one* of the agent's values. It may also reflect numerous arbitrary causal influences (mediated or unmediated) that the agent would not recognize as justificatory; but that is not enough to disqualify it as agentive."

If meeting this requirement is supposed to be sufficient for agency, difficulty ensues.[5] Say I value both gustatory pleasure and health, but value gustatory pleasure far less. Say next that someone slips me a science fiction-y medication that briefly amplifies, in some rationally arbitrary fashion, the value I place on gustatory pleasure, and I end up elbow deep in a tray of Winslow's salt caramel donuts. On **Ward & Machery**'s amendment, I'm responsible, because my snarfing expresses at least one of my values. But manipulation of this kind is a classic responsibility negating excuse. (More on manipulation below.) The point generalizes: even manifestly non-agential behavior may express one (or more) of an actor's values. While Ward & Machery's amendment stems the unwelcome proliferation of defeaters, it may succeed too well, countenancing the unwelcome proliferation of agency.

My own attempt to curb the population of defeaters is to say defeaters obtain when the actor would be "unwilling to cite in defense of her behavior the factors figuring in the most perspicuous psychological explanation of her behavior" (précis, p. 1). What's meant to do the needed work is the admittedly vague "most perspicuous psychological explanation." The most perspicuous explanations won't include irrelevant or extraneous information: many rationally arbitrary causal factors that do not undermine agency, like my parents' first meeting, are in this way excluded. Also doing some work is the requirement that the explanations be "psychological"; when we ask about motives and reasons, we're not asking about distal natural causes like the big bang, so these candidate defeaters are excluded. Is my way of culling defeaters ad hoc or good sense?

Both. Assessing explanatory relevance proceeds on a case-by-case basis, and there's unlikely to be anything tidy and general to say about whether something implicated in a behavior destabilizes agency, or is merely an agentially irrelevant feature of that behavior's causal history. Perhaps this amounts to an on-the-fly "sniff test," but I'm guessing such appeals are inevitably part of assessing explanations, in as much as explanation has pragmatic goals like producing understanding.[6] There may be some biting of bullets here, where I have to recognize defeaters where I'd rather not. That is, I may get a bit less agency than I wish. But without further elaboration, **Ward & Machery**'s proposal seems to leave us with agency where we shouldn't have it, so I'll stick to my approach, messy as it is.

R1.3. Skepticism and practical utility

Speaking of explanations, **Lambert & Dennett** demand one of me: if rationally arbitrary influences on behavior are as robust a phenomenon as I suppose, "how do we manage to hold it all together?" People do pretty well at things like interpretation, prediction, and coordination – and more generally, just getting by – but this practical success would be miraculous, the argument goes, if agency were regularly undermined by unexpected "goofy" influences.

Here, it is useful to distinguish *rigid* and *fluid* contexts. Rigid contexts are highly constrained (whether implicitly or explicitly), and any exercises of agency in these domains are likely to be less easily destabilized. Fluid contexts are less constrained, and present more latitude for defeaters. Under ordinary conditions, driving is a fairly rigid context: not often do goofy influences have you speeding the wrong direction on the interstate. But there are many (explicit) protections in place, like routing and signage, to prevent such mishaps. Similarly, I doubt there's some goofy manipulation of the sort found in the psychology literature that could surreptitiously induce me to lecture naked, in face of the countervailing (often implicit) social expectations. Once more, this looks like a rigid context, and there's doesn't seem to be much question about defeaters.

Now think of the complex, emotionally freighted contexts I'm calling fluid, such as romantic relationships, political preferences, and career choices. In these cases, which are often of great practical interest, there are serious questions about defeaters (pp. 75–76). Regarding **Lambert & Dennett**'s question, I'm inclined to think that much of our practical success in endeavors like prediction and coordination can be attributed to the behavioral regularity enforced by rigid contexts, while the troubling failures of agency, I'd wager, more often occur in fluid contexts, where there are weaker forces for keeping people on the rails. If so, we have the paradoxical result that defeaters are less likely, and the exercise of agency more likely, where behavior is more constrained. Given that my theory of agency "positively celebrates constraint" (p. 12), this air of paradox is not untoward. It may also be that the rigid/fluid distinction provides some traction on the epistemological problem: we may more confidently attribute agency in certain rigid contexts.

Hirstein & Sifferd offer another sort of explanatory demand: "Doris's view amounts to saying that the entire upper level that has been designed into our brains,

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

including the executive processes and consciousness itself, is of little use or import." I'm generally leery of arguments about what it makes sense for Mother Nature to do, but if it were the case that my theory implies She designed-in an expensive system without assigning it meaningful work, I agree I'd have some explaining to do. I didn't, as Hirstein & Sifferd observe, say much about brain science (pp. 88–89), but in my dialogic theory the collaborative exchange of rationalizations makes plenty of toil for the "upper level" (as, presumably, do those occasions where behavior conforms to a traditional reflectivist model). Additionally, the brain does many things that are not in the service of morally responsible agency, and any number of these might employ the upper level. I'm inclined to think, with theorists like Mercier and Sperber (2009) and von Hippel and Trivers (2011) that nature didn't design the fancy human brain with heavy emphasis on accurate self-awareness, but that doesn't mean there's nothing for it to do.

Hammond leverages developmental psychology, which I left largely untapped, into a suggestive defense of reflection. His central example is a study where Piaget (1974/1976) had children and adults model crawling. Not everyone hit on the correct "X" pattern; for example, some subjects modeled crawling on an "N" pattern. But when subsequently asked to crawl, some of the mistaken modelers crawled in accordance with their erroneous model. As Hammond has it, "our reflections can shape our actions, *even when these reflections are inaccurate representations of the state of the world,*" and these "subsequently reorganized actions may create a world that more closely resembles what was in error."

I take a similar line (pp. 135–37, 143–45): the fact can be the child of the fiction, in something like the way **Hammond** supposes. But now, as a self-styled anti-reflectivist, I have a branding issue. For what Hammond is proposing, in effect, is reflectivism minus the accuracy corollary. While I'm apt to caution that much agential behavior is unreflective (p. 69), perhaps my view ought be characterized as a kind of reflectivism, in so far as I recognize the kind of process Hammond describes. On the other hand, jettisoning the accuracy corollary might cause the reflectivist branding problems of her own: the Oracle said *Know thyself*, not *Reflect inaccurately about thyself, and structure future behavior with these inaccuracies*. A view like mine (and I think Hammond's), which centrally features *inaccurate* self-awareness, betrays the spirit of traditional reflectivisms.

Collerton & Perry defend the practical efficacy of reflection by means of an analogy with vision. Vision is prey to many distortions and inaccuracies, but is nevertheless enormously helpful for getting us along; likewise, Collerton & Perry say, for reflection, which needn't be perfect to be useful. I'm inclined to agree (pp. 129–33), but I'd caution against supposing that our success in getting along is evidence for our successful exercise of agency. After all, many species that presumably lack morally responsible agency, like the pathetically robotical *Sphex* wasp made philosophically famous by Dennett (1984, pp. 10–11), manage just fine (when not made examples of by meddling naturalists),[7] and it's not clear that we wouldn't be able to do the same if we lacked morally responsible agency – as indeed some philosophers (e.g., Pereboom 2014) have argued.

The analogy with vision actually presses this point. Noting the practical value of vision does not necessarily answer skepticism directed at the deliverances of the senses; the skeptic will be quick to grant practical efficacy, while continuing to question the possibility of knowledge. Indeed, comparatively simple animals successfully navigate their world, but epistemologists may hesitate to ascribe full-blooded knowledge to such organisms (e.g., Sosa 1991, p. 240). Likewise, the agency skeptic isn't doubting that human beings get by, she's doubting that they get by while exercising agency. And this doubt, she thinks, doesn't get allayed by insisting on our pragmatic success.

R1.4. Skepticism and individual differences

Several commentators suggest I should pay more attention to individual differences. While I've sometimes criticized character and personality theory, I've always acknowledged the importance of individual differences (e.g., Doris 2002, pp. 25–26), and I participate in empirical work investigating them (Bollich et al. 2016; Cameron et al. 2013;[8] Mooijman et al., forthcoming). Although individual differences were not a primary focus of *Talking to Our Selves*, I assumed throughout that there are individual differences with respect to agency (e.g., pp. 34–35, 39–41, 48, 156, 162).

It is an interesting, and underexplored, question what exactly those differences might be; one major obstacle is operationalizing amorphous and contested notions of agency for empirical work. **Lambert & Dennett** suggest there are likely individual differences in susceptibility to the goofy influences that may function as defeaters; if so, there should by individual differences in the extent to which people exercise agency. For example, there's considerable research on individual differences in suggestibility (e. g., Frost et al. 2013; Marotta et al. 2016). Intuitively, the less suggestible might be less susceptible to goofy influences, so maybe the less suggestible are, all else equal, better able to exercise agency. This, it seems to me, has the makings of a worthwhile research program integrating moral psychology and personality psychology.

Niemi & Graham argue that individual differences affect not only the exercise of agency, but also the attribution of agency. Work by Graham and his colleagues (Graham et al. 2009; Graham et al. 2011) on *moral foundations* suggests that different people may adopt very different moral perspectives; for example, conservatives may favor *binding* values like loyalty and purity, which support group solidarity, while liberals may favor *individualizing* values like harm prevention and fairness, which focus on the protection of individual persons. Niemi and Young (2016) have shown that the differing values may be implicated in differing perspectives on responsibility; for example, those higher in binding values may attribute more responsibility to the victim of sexual assault. This result might be thought to generate another sort of skeptical challenge: if variation in responsibility attribution is attributable to foundational differences in values, perhaps we cannot expect agreement on responsibility attributions between people embodying these differing evaluative perspectives. And if there is such "fundamental" evaluative disagreement (Doris & Plakias 2007), perhaps we cannot converge on objective assessments of responsibility.[9]

While consideration of individual differences may be thought to exacerbate the skeptical difficulty, **Taylor**

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

thinks it may instead suggest a solution: "the development of expertise is a widely applicable self-regulative strategy" so "people should develop expertise in certain domains if they want to consistently express their values." It's a truism of the human performance literature that where there's expertise, there are individual differences, at least in complex domains (Gobet & Campitelli 2007, p. 159; Howe et al. 1998, pp. 399–400), so if exercising agency involves a kind of expertise, some folks should be better agents than others. This may sound a bit undemocratic, but in as much as the development of expertise is sensitive to practice (e. g., Ericsson 2014), perhaps the less agential may become more so, if they put in the work.

Taylor's suggestion is intriguing, not least because expert performance is, oftentimes, not so much undermined by automatic processing as supported by it (Christensen et al. 2016), and the performance of experts often seems appropriately agential. One of Taylor's examples is a study where nurses were less subject to the "group effects" that mute helping behavior when other bystanders are present (Cramer et al. 1988). Arguably, the nurses' training made them less susceptible to defeaters, and more agential, in the performance of their profession. Expertise, we might say, involves a kind of educated automaticity: if you want to exercise agency in an area, and squelch lurking defeaters, develop expertise.

Unfortunately, the development of expertise is not so well understood as we might wish, regarding such questions as the relative contributions of talent and practice (for an overview, see Doris, in preparation). Additionally, expertise is domain limited (Chi 2006): given the large amount of practice required, and the unequal distribution of domain-relevant talents, people are seldom truly expert in multiple complex domains. Now, **Taylor**'s intriguing proposal has generated another intriguing proposal. Perhaps, if people are only expert in a few (or fewer) domains, people only exercise agency in a few narrowly circumscribed areas of their life: maybe I exercise agency as an academic, but not as an athlete. If so, Taylor's solution to the skeptical problem may suggest that for most people, the scope of agency is sharply curtailed.

Fowers et al. also urge attention to individual differences as a way of defanging skepticism. As I say, I welcome research on individual differences, but I'd caution against proceeding in terms of virtue, as they propose: "Behavioral research on virtue and character is just getting under way . . ., but the expectation is that virtue and character will directly reduce defeaters' influence as well as moderate defeater effects." (After critiquing the philosophical method, Fowers et al. here appeal not to empirical evidence, but to the philosophical authority of Aristotle.)

Apparently, Fowers et al. wish to address the skepticism about traditional conceptions of character I developed in an earlier book (Doris 2002). Although I have thoughts about how the literature subsequent to that book is developing (Doris, forthcoming), in the book that is the topic of the present discussion, I explicitly declined to revisit the issue (pp. 14–16). The character skepticism debate in philosophy (like the person-situation debate in psychology that inspired it) is substantially a debate about cross-situational behavioral consistency – the skeptics contend behavior is much less consistent than traditional notions of character would have us predict. But that issue is orthogonal to the issue of agency. To see this, consider some non-human organisms that exhibit limited behavioral variation, like sea anemones. An organism could be *perfectly* consistent, and not be an agent; in my frame, a person could be both perfectly consistent and invariably subject to defeaters.

Moreover, linking virtue (or moral probity more generally) and agency has unpalatable results (p. 16): if virtue is what facilitates agency, it would appear that the vicious are unlikely to exercise agency, which flies in the face of the apparently viable practice of holding less-than-virtuous people responsible for their bad deeds. Or so I argue (pp. 156–59). You needn't be convinced. But if you link virtue and agency, you must either (1) develop distinct accounts of responsibility for wrongdoing and rightdoing, or (2) contend that non-virtuous people occupy mitigating or exempting conditions and eschew, contrary to ordinary practice, holding them responsible. Because neither of these options is completely appealing, you'd want to offer considerable argument. Fowers et al. do not attempt to do so.

R2. The anti-skeptical theory

I've reviewed the skeptical challenge and considered some avenues of response that were not featured in *Talking to Our Selves*. I now turn now to the anti-skeptical response I did feature, and consider some of my commentators' thoughts on that.

R2.1. Do we need a theory of agency?

Patrzyk decries the "moralistic appeal of agency attribution," and any attempt to vindicate existing practices of responsibility attribution. Anti-skeptical approaches to moral responsibility, he contends, are "rescue missions," which can only be explained by theorists' "need to devise a theory allowing them to justify their retributive instincts and take credit for what they do not deserve." Existing responsibility practices, for him, are "objectively unfair, creating a situation in which blame judgments depend on factors that cannot be controlled by the actor."

While **Patrzyk** finds my moralizing distasteful, he himself is moralizing, because he finds non-skeptical thinking on morality "objectively unfair." I suspect that charges of moralism themselves reliably lapse into moralizing – the more so when deploying markers of certitude like "objectively." Given what he says, perhaps Patrzyk should be embarrassed by his moralizing, but I readily admit that my own theorizing is morally freighted.

This is inevitable, in as much as notions like agency and responsibility are "thick," and simultaneously bear normative and evaluative commitments (pp. 14, 195–96). Thus, asserting that people sometimes exercise morally responsible agency, as I do, is an ethical exercise, as is *denying* that they do so; the former is to say that people are due a certain kind of moral regard, the latter is to say they are not (as indeed **Patrzyk**'s "credit for what they do not deserve" makes obvious). For example, I'm of the altogether familiar conviction that people deserve a different kind of regard for their good deeds than for their good looks, and I'd mark this divide with the language of agency. Patrzyk may either reject this conviction (as philosophers with skeptical leanings about responsibility, such as Smart [1961], are

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

perhaps inclined), or develop a way of grounding it that does not appeal to agency. In doing either, he would be engaging in a recognizably moral enterprise; in debating agency and responsibility, there's no escape from "moralism."

R2.2. Reactive attitudes

I anchor my approach in the "reactive attitudes," but as **Viciana, Gaitán, & Aguiar** (**Viciana et al.**) observe, my account here is "importantly undertheorized." Strawson (1962), who initially proposed understanding responsibility *via* the reactive attitudes, was himself more suggestive than systematic, and I was perhaps overly content to emulate him in this respect. I supposed that reactive attitudes like anger and admiration are "symptoms" of morally responsible agency (p. 24), while Viciana et al. assert that reactive attitudes "can sometimes be apt even if they are not tracking morally responsible agency." They suggest, for example, that we ought be outraged at perpetrators of wartime atrocities, even if the conditions of war often destabilize miscreants' agency (as Doris & Murphy [2007] also argue).

I take the point: perhaps the reactive attitudes, like the assignment of criminal responsibility, are sometimes governed by something like strict liability, and appropriately target perpetrators of non-agential deeds. Conversely, there may be cases of responsible agency that don't prompt the reactive attitudes; for example, some agential behaviors may be too inconsequential to provoke the kind of emotional responses, like anger, associated with the reactive attitudes (possibly, this departs what I said in the book, on p. 24).

These are important points, but I doubt they vitiate my approach. I propose treating reactive attitudes as a sort of heuristic – a way of loosely delineating the contours of responsibility practices (*plural* practices, because as **Viciana et al.** note, not all responsibility systems are the same). Just as it is often the case that no one symptom is necessary or sufficient (and still less, necessary *and* sufficient) for diagnosing disease, the presence of reactive attitudes is not necessary or sufficient evidence that agency is at issue. Still, if one wants to look for responsibility attributions, and determine whether those attributions are legitimate, one could do far worse than looking for the presence of the reactive attitudes, together with the kind of justifications people offer when they experience and express them. The reactive attitudes are, as I treat them, a beginning – a way into a theory of morally responsible agency, not the theory itself (p. 23).

R2.3. Agency and values

On the theory I propose, exercises of agency are associated with expressions of value: when my declining a donut expresses the value I place on my health, I exercise morally responsible agency. **Dunning** wonders why I privilege values, among the various psychological states and processes that structure behavior. My thought is that the relevant psychological states must have a certain *authority*. For example, lots of desires I have are unimportant, or even repellent, to me, and many others may be fleeting and infirm; they're not the sorts of things around which it makes sense to structure my behavior – and still less, my life. My account of values attempts to identify "desiderative complexes" (p. 183) that are fit to play this structuring role: "values are associated with desires that exhibit some degree of strength, duration, ultimacy, and non-fungibility, while playing a determinative-justificatory role in planning" (p. 28). Such desires, we might say, are worth taking seriously: They have *normative heft*, or *rational authority*. I agree with Dunning that I might have gone more pluralist here, and not limited my account to values; perhaps it's possible to "upgrade" other psychological states, like attitudes, in the way I attempted for desires, so that these states can also be seen as integral to agency. Nevertheless, I'd want this pluralism to reserve a large role for values, because they are plausibly supposed to bear the needed normative heft.

Maibom worries that on my account, "there may be *no* causal chain from your values to the expression of them"; if no causal relation is required, even things I didn't do might count as exercises of my agency (like someone I've never met bringing about an outcome I value after my death). For this reason, I distinguished *conforming to* a value and *expressing* a value, arguing that expression requires more than mere conformity (p. 70). Part of the something more, I said, is a causal relation: for expression, "that the actor holds the value should be causally implicated in her undertaking a behavior suited to realize the value" (p. 135). But a causal relation will not be enough, because it might be fortuitous or fluky (p. 25): it shouldn't count as an exercise of agency if the value I place on health causes me to faint in horror rather than lighting up when you offer me a smoke as I'm trying to quit. My fix was to propose that the expression relation is manifest in intentional behavior, which I construed as goal-directed behavior (pp. 25–26); thus, "a behavior expresses a value . . . when that behavior is guided by a value-relevant goal." This helps exclude the fortuitous and fluky from the offices of agency.

In addition to worrying that my theory makes for too much morally responsible agency, **Maibom** worries that it makes for too little (as I myself do [p. 70]). On my view, she thinks, "people are rarely, if ever, responsible for wrongdoing," for they are "unlikely to have put desires in the driving seat that are the sorts of values that we see expressed in wrongdoing." If putting in "the driver's seat" involves conscious decision, I agree that many, perhaps most, wrongdoers do not act as Milton's Satan ("Evil, be thou my Good"). But crucially, on my view people do not need to consciously entertain the values on which they act to exercise agency; indeed, they might be quite unaware of having the value expressed in these exercises (pp. 27–28, 160–61). Moreover, a value does not need to be intrinsically bad to be expressed in wrongdoing, which will often depend on context: valuing sexual gratification is not wicked, but expressing that value while betraying an explicitly monogamous relationship can be. And of course, in many cases there will be more than one value in play, and sometimes wrongdoing stems less from badness of one's values than from the way one integrates or weights them (cf. p. 162). Hopefully, then, my theory makes for neither too little agency nor too much.

Murray argues that the valuational theory cannot properly account for inadvertent "slips," like neglecting to turn off the stove and starting a fire, because many slips are omissions for which people are appropriately attributed responsibility, despite the fact that these omissions may

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

"not express a relevant subset of the agent's values." Like **Maibom**, Murray worries that my approach is overly exculpating: The unlucky cook is properly held responsible, he thinks, despite not valuing house fires.

Many slips, it seems to me, admit of valuational explanations (some Freudian slips, for example). Perhaps the forgetful cook insufficiently values prudential precautions like double checking to see that the stove is off, and the slip can be characterized by the valuational theory as responsible negligence, despite it being the case that a house fire does not reflect what **Murray** calls the actor's "overall balance of preferences." In other cases, perhaps there is no relevant value *anywhere in the vicinity* of the slip. But here, I don't find myself much inclined to attribute responsibility (p. 155), as in the unfathomably tragic cases where ordinarily conscientious parents leave their young child in the car to die of hyperthermia (for more on slips, see Amaya & Doris 2015).

Maybe you disagree with my lenience here; **Murray** asserts that "we regularly hold people responsible for their slips" and thinks "we should prefer theories of responsible agency that preserve this part of our practices" – a preference he thinks my theoretical conservativism requires (p. 158). I'm not sure what the contours of everyday practices regarding slips are, but I'm betting they're pretty variable; while I don't want to lean too hard on the law, parents implicated in fatal vehicular hyperthermia accidents are prosecuted around half the time (Collins 2006, pp. 807 n. 2, 825). My guess is that people are sometimes held responsible for slips and sometimes not; "I just forgot" can, after all, sometimes be an acceptable excuse. (That it isn't acceptable for hyperthermia accidents is probably attributable more to the horrific outcome than the psychological inaccuracy of "I just forgot.") So I'm betting the valuational account can manage the intuitive verdict for a tolerable percentage of slips. But it wouldn't much trouble me if it can't. My approach is "revisionary in some regards and conservative in others" (p. 158), allowing that there is "endless room for improvement" in habits of responsibility attribution. If Murray's speculation that people are routinely held responsible for slips is right, I'm inclined to say the practice overreaches and perhaps ought be changed, a revisionary suggestion that is entirely compatible with my approach.

Vargas thinks the valuational account is prey to "manipulation cases": a manipulated person might be expressing their values (indeed, she might be manipulated *into* expressing her values), but manipulation is a standard excuse, so my approach holds people responsible who are not. My response is dilemmatic (cf. pp. 31–32): either the manipulation effects impairments in the exercise of capacities requisite for responsibility, or it does not. If it does effect responsibility relevant impairments, the manipulation is responsibility negating (or mitigating), and I can say the intuitive thing. If it does not, I must bite the bullet and hold the manipulation victim responsible. To soften the bite, I'd contend that cases of manipulation without responsibility relevant impairments are empirically unlikely, so the awkwardness will be unusual.

But, **Vargas** argues, there is a phenomenon making similar difficulty for me that is not only empirically likely, but also in fact commonplace: "adaptive preferences," where people's motives and desires are shaped by oppression. Sexism, racism, or other patterns of prejudice, for example, might instill self-abnegating desires (Westlund 2003). According to Vargas, such desires can be associated with values, which means my account apparently has the victims responsible when their conduct expresses the values that result from oppression.

This is an exceedingly sensitive issue. While it may seem hard-hearted to hold victims of oppression responsible for value driven behavior attributable to their oppression, it may seem paternalistic not to hold them responsible for acting in accordance with their values – the more so, where one person's *culture* is another's *oppression*. For one of many examples intimating the difficulty, consider debates about Sharia and feminism: Is observing certain traditional religious practices compatible with progressive notions of self-determination and autonomy?[10]

Let's suppose, however, that there are clear cases of adaptive preferences where the subject should not be held responsible for the associated behavior. If these preferences should be thought of as values, the valuational theory apparently gets the wrong result, holding the actors responsible for behavior expressing these values. However, I question whether the desires at issue have the status of values. The prisoner might (in some sense) desire the disgusting institutional food without valuing it, just as the torture victim might (in some sense) desire denouncing his country, without valuing this performance. Remember that on my view (p. 28), values are associated with ultimate desires that figure in justification and planning; the adaptive desires of the prisoner and torture victim are not ultimate, but instrumental, in the service of their survival. Furthermore, it may be that people repudiate their adaptive preferences even while in the oppressive conditions, and would not appeal to them in justification and planning.

So I'm inclined to think that adaptive preferences will often fall short of values. Where they are aptly characterized as values, there's still the question of whether the associated oppression results in responsibility negating or mitigating disabilities; for example, oppression may result in exculpating ignorance. But when that is not the case, and the adaptive preference is properly thought of as a value, my account attributes responsibility for behaviors expressing that value. This may be an unhappy result. It is a difficulty common to the many "currentist" theories of responsibility that, like mine (pp. 30–32), deny that historical considerations figure directly into assessments of responsibility (as opposed to being *indirectly* accounted for by assessment of the resulting current states). But if I'm right, such embarrassment will be acceptably scarce.

R2.4. Evolutionary theory

Beal & Rochat observe that I make little use of evolutionary theory, which, they think, means I join the story too late: Instead of the negotiation of moral responsibility, I should have been speaking of a "renegotiation," for human organisms have preferences even in the womb. As I'd put it, *biology constrains agency* (cf. p. 195): even in a world of massive cultural diversity, not all forms of life are available, and whatever agency is open to us, it will be highly canalized. In this spirit, Beal & Rochat suggest that I ought better attend to "primordial" features of our natural and material worlds, which also serve to structure and sustain the self: for example, in attempting to

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

understand the tragedy befalling displaced peoples, we must recognize that in addition to social and cultural rupture, the loss of land, animals, and other elements of the natural world is itself catastrophic. I'm quite happy to adopt Beal & Rochat's observations, which I think can enrich my account; as I said, I'd be perturbed if my approach were incompatible with the best evolutionary theory, but I'm optimistic this isn't the case (p. 145).

R2.5. Culture

Franks & Voyer and **Dunning** observe that cultures vary not only in their values, but also in their understandings of agency. In the book I tentatively concluded (pp. 192–96), from cross-cultural research on the locus of control, that the notion of agency I had in mind was perhaps more widespread than the most provocative readings of cultural psychology, like those adverted to by Franks & Voyer and Dunning, would have us suppose.[11] But I also took the view that I'd not much fret if the scope of my topic was more parochial than I supposed. WEIRD people –those who are Western, educated, industrialized, rich, and democratic (Henrich et al. 2010) – aren't the most numerous of the world's peoples, but they are numerous enough to be worth thinking about, and they were, I suppose, the (mostly tacit) focus of my book. Identifying theoretically perspicuous underpinnings for a – revisably – viable cultural practice is honest work, I thought, even if that practice is parochial.

Perhaps this was overly complacent. **Dunning** suggests, and I take it **Franks & Voyer** would agree, that I overemphasize a Western, *disjoint,* conceptions of agency, where agency involves individuals "imposing their will on the external world." He commends more attention to *conjoint* conceptions of agency, as found in some non-Western cultures, where "people strive to harmonize their actions with outside forces and constraints, usually social ones." Maybe the notion of me as an individual having my *own,* independent, values resonates much less in conjoint cultures than it does in disjoint cultures. Maybe in conjoint cultures the primary notion of agency is group agency, rather than the individual agency that seems to be everywhere celebrated in the West. My interest in the book was individual agency (p. 169), but if conjoint, group, agency is prominent across the cultures of the world, it would behoove me to explore extending my approach to group agency; my guess is that my collaborativism is well suited for this task, though I can't take that up here.

Another possibility deserving further attention is that cross-cultural diversity amplifies the skeptical challenge. **Uleman, Granot, & Shimizu** argue that the role of cultural and contextual influences on attribution of responsibility means there is unlikely to be an "objectively 'correct'" or "god's-eye" view of moral responsibility grounding univocal answers to questions of who is responsible. Shimizu et al. (2017) propose that most important cultural differences occur through automatic processes; taken together with Uleman and colleagues' work on spontaneous social inference (e.g., Uleman et al. 2012), this leads me to expect that many moral judgments, including responsibility attributions, will tend to be culturally determined, automatic, and unreflective. The attribution of responsibility may itself be subject to defeaters, no less than are the behaviors that are (together with the actor) the targets of responsibility attribution.

That is, people's attributions may be influenced by factors they would not, were they aware of them, invoke as justifications for their attribution. And where people are unaware of these influences, they may be unable to resist them. Now the skeptical difficulty affixes at a second place: not to the *exercise* of morally responsible agency, but to the *attribution* of morally responsible agency. If an attribution of responsibility is to be warranted, the presence of defeaters must twice be ruled out, once for action, and once for attribution. Then the challenge of developing the sort of epistemically robust collaborations I propose in response to skepticism is even more challenging than I'd imagined.

R2.6 Collaborativism: Pitfalls

Pe-Curto et al. raise what they call "the problem of bad company," a difficulty also marked by **Lambert & Dennett.** As **Couchman, Birster, & Coutinho** (**Couchman et al.**) put it, not all participants in the "dialog of self" I envisage will be "playing nice": people's values might be "hijacked" by manipulative social interactions, and even where participants have the best of intentions, agential dialog may be undermined by biases like anchoring effects.

Perhaps sociality is as likely to impair as promote agency: sociality plus totalitarianism, for example, might vitiate agency rather than facilitate it. **Pe-Curto et al.** sharpen this difficulty with an epistemological challenge: "If we are so entangled in our milieu for cognition, agency, and our unity as selves, we appear badly placed to tell whether we should embrace or resist its influence."

Pe-Curto et al.'s response is reflectivist-individualist: "Should we be in bad company, we might need *reflective, individualist* humans sufficiently in touch with their values, and so able to disentangle themselves from social influence. If we may be so bold to suggest it, we might need humans of character." Relatedly, **Couchman et al.** assert that "the highest ethical standard in [Doris's] system ought to be the process of increasing metacognition – the ability to self-regulate (to beat defeaters) and to avoid biases and hostile narratives (or meta-defeaters, if you will)."

As I see it, both of these solutions are troubled by the vagaries of individual reflection I documented throughout *Talking to Our Selves*; **Pe-Curto et al.** and **Couchman et al.** locate the solution where I've located the problem. Regarding Pe-Curto et al.'s appeal to character as the facilitator of reflection, I'd want to know more about what aspects of character get the call in this role, and how they are developed, and I'd also appeal to my work on frailty of character and the uncertainty of moral education (Doris 2002; forthcoming; in preparation). Regarding Couchman et al.'s appeal to metacognition, I don't deny that metacognition has a role in agency – I say the same for reflection generally (pp. 74, 171–77). Nevertheless, their proposal is subject to the very difficulties they themselves raise: why think metacognition less liable to bias and hostile manipulation than cognition *simpliciter*? (If the solution is to reflectively monitor metacognitions, what makes these meta-metacognitions immune? And so on.) This suggests that Couchman et al. have not ameliorated, but instead relocated, the skeptical difficulty.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

But **Couchman et al.** also suggest the kind of ameliorative strategy I favor, one that trades more on external scaffolding than internal cognition. Knowing the vagaries of memory, an organization might provide minutes of a meeting to attendees. Concerned about the possibility of implicit bias, a search committee might appoint a "diversity officer" to ensure that all files from underrepresented groups receive full and fair consideration. Increasing the accuracy of metacognition can have a place in my approach, but it won't be the "highest ethical standard," because other processes, including the cultural and institutional, will be at least as important to agency. Of course, what's required is the "right kind" of relationships and institutions, and I didn't do nearly as much as needs be done in specifying what the right kinds are (but see pp. 119–23). To go further, I'm guessing, there will need to be much closer connections between moral psychology and such disciplines as political science than are currently evident in the literature.

None of what I've said excludes altogether a role for reflection; just as there aren't tight entailments between reflectivism and individualism, collaborativism does not entail the rejection of reflectivism (p. 110). Reflection has a role in human life, and a role in agency. The trick is to say something about when and how. **Collerton & Perry** offer a rich example: the distressing hallucinations associated with some forms of mental illness may be improved by "a joint therapist-client investigation of the reality of experiences." This much, as they note, is congenial to my collaborativism, but they also suggest the process involves *accurate* reflection, because "insight" may contribute to amelioration. Then defeaters may sometimes be countered by accurate reflection, and people may sometimes better exercise agency in this way. The clinical treatment of hallucinations is a limited context, but I don't deny that there may be others; that such contexts are prominent enough in human life to satisfy the reflectivist, I am inclined to doubt.

It's also worth noting that reflection may have importance beyond facilitating the exercise of agency. **Franks & Voyer** contend that my focus on "revealed" agency – agency as manifested in (patterns of) overt behavior – neglects *experiences* of agency, which "figure significantly in people's own normative explanations and justifications, and connect directly to the sense of self." This seems right to me: that I think of myself as an agent, rather than, say, a puppet, has a lot to do with how I think and feel about myself, and how others do. An illustration of this I find especially compelling is abnormal experiences of control in mental illness, such as the atrophied perceptions of control associated with schizophrenia (p. 134); the significance (and misfortune) of this condition is not limited to any associated impairments of agential behavior.

R2.7. Collaborativism: Processes

Bonicalzi & Gallotti observe that more needs be said about the "mechanics of collaborativism." A mechanism they suggest is *alignment* – the development of shared understandings, such as publicly available and validated moral norms, through social interaction. The internalization of these norms may help answer an important question for collaborativism: why should people be moved to justify themselves to each other, and engage in the practice of rationalization at all? Part of the answer, evidently, is that they hold themselves, to some degree, to a set of shared norms, and so feel motivated to live up to the norms, and explain themselves when they do not. (This account of course requires a story about how the norms get internalized [e.g., Sripada & Stich 2006].) So if Bonicalzi & Gallotti are right, we've got a crucial piece of the needed mechanics: a story about why people are motivated to collaborate in the first place.

I'm pleased that **Niemi & Graham** considered my account of the self from the book's last chapter, which seemed to draw less attention than other parts of my argument, but is, I think, crucial to understanding collaborative agency. My concern there was destabilizations of self and agency: I contended (pp. 181–86) that cultural devastation might disrupt the personal continuity required for temporally extended "diachronic" (pp. 163–64) exercises of agency associated with the major life projects, like those involving work and family, that imbue human life with meaning.

Niemi & Graham suggest a process by which this might occur: the depersonalization associated with trauma. Work such as Nizzi and Niemi's (in preparation) suggests that trauma survivors may experience a sense of self rupture and foreshortened future: their pre-trauma self seems to them destroyed, and it is unclear to them how, or if, their post-trauma self may go forward. This strikingly resonates with the remark attributed to the Crow Chief Plenty Coups, who says of the destruction of Crow traditions, "after this, nothing happened" (pp. 180–81). I didn't notice the connection between cultural devastation and trauma research before Niemi & Graham's urgings, and I take their suggestion that my theory has testable empirical implications regarding the experience of self, patterns of moral judgment, and commission of harm. I also welcome their proposal for future directions: collaborations between clinical psychology and philosophy for understanding and addressing trauma. Philosophical theorizing about agency and the self may help us to understand trauma, while understanding the clinical processes by which trauma may be healed can help us to understand the mechanisms that develop and sustain agency and the self.

Hechler & Kessler indicate another important direction for enriching collaborativism. I understood the negotiations characteristic of agency and its attribution in terms of simple dyads, but as they say, agency attributions are often produced by multiple observers, who "validate their perceptions and beliefs with reference to their fellow group members." Attribution is not limited to straightforward actor-observer pairings: there's also actor-observer**s**, actor**s**-observer, actor**s**-observer**s**. Also, because observers will often be active rather than passive observers, many attributions may be better described in terms of actor(s)-actor(s) dynamics (that is, the actor-observer distinction is unstable). Also, as Hechler & Kessler say, attribution may express values (cf. **Niemi & Graham**), which means that the attribution of agency may also be an exercise of agency. Finally, the attribution of agency may facilitate the exercise of agency; agency may involve collaborations among groups who are exercising their own agency while facilitating and constraining the agency of other groups, who are themselves doing the same. Collaboration, then, is likely to be far more complex than my programmatic depictions intimate.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Oftentimes, these complex social dynamics will not be well described as collaborative; as **Hechler & Kessler** observe, the parties may stand in varying relationships from cooperative to adversarial. This, however, won't always undermine agency, as **Mercier** shows in augmenting collaborativism with the *interactionist* account of reasoning he developed with Sperber (Mercier & Sperber 2017). On their view, reason "would have evolved chiefly to serve two related functions, which are both social": (1) justifying our actions, and evaluating other's justifications, for the purpose of social assessment, and (2) arguing for our own beliefs and evaluating the arguments of others, for the purpose of facilitating communication.

At present, the keyword is "arguing." I had in mind cooperative collaborations, where participants have substantially overlapping interests, as in my favored example of romantic relationships. But **Mercier** says, "reasons are most helpful when full collaboration cannot be expected"; the call for justification arises more when parties disagree than when they agree. In so far as my justifications structure my behavior in ways that express my values, it appears to follow that contention, no less than collaboration, may facilitate agency. While we get by with the help of our friends, we may also need the help of our frenemies: Bad Company isn't always bad for agency.

For a fully baked rendering of collaborativism, we require an understanding of "reasoning" more developed than my rather programmatic account (pp. 43–44; 104–106). **Mercier** counsels against equating reasoning, as I was tempted to do (p. 50), with "System 2" effortful, analytic processing, because "finding and evaluating reasons is, in most cases, quasi-effortless and automatic." I found conflicts between the dumb automatic and smart analytic to be most trenchant (pp. 69–70); but if Mercier is right, many conflicts between reason and unreason may be conflicts *within* automatic processing. This might, as he suggests, mean my critique of reflectivism is "too generous," because the role of reflection might be limited even within the class of cognitive activities appropriately considered reasoning.

That said, we're not without anti-skeptical resources. **Zinken & Reddy**, with a nice turn of phrase, suggest that in addition to defeaters, there are *supporters*, "causes of our behavior that we are often unaware of, but that would make good-enough reasons for our actions, were we made aware of them." They launch this suggestion with intriguing work in linguistics on the selection of interrogatives – *Can you pass me a plate?* – or imperatives – *Pass me a plate* – in making a request (e. g., Zinken & Ogiermann 2013). Evidently, the selection depends on how intrusive the request is: if passing a plate is *continuous* with what you are doing – maybe you are stacking dishes – my using the imperative is fine, but if it is *discontinuous* – maybe you are stirring the sauce – an interrogative is required. One lesson for my project is that linguistics, which I did not consider in the book, offers materials for helping us better understand collaborativism: language shapes social interactions, so understanding the details of this shaping might help us better understand how sociality facilitates the exercise of agency.

Additionally, these findings from linguistics can enrich the understanding of defeaters. Presumably, the grammatical form of a request – for everyday requests involving plates, if not for life-changing requests like proposals of marriage – is very often selected unconsciously. But if you were made aware that your selection was based on assessments of continuity, would you regard this as a reason? Presumably, you wouldn't spontaneously offer up such a reason; Zinken's (2016) subjects justified their selection by appeal to politeness, or said it was arbitrary. If you were offered the scientifically substantiated "continuity explanation" would you take it on board as a "good enough reason"? And if you took it on board as a reason, would that suggest your behavior was agential?

I suspect you didn't know about the explanation, and why it makes your behavior make sense, on previous occasions when you've made a request – it's a cutting-edge bit of scientific discovery and theorizing, that, if you are like me, you're just learning about. If you would accept the scientific account as your reason, when it was explained to you, it apparently makes a promising candidate for agency on my view (pp. 27–28), which requires only counterfactual acceptance (a necessary amendment, because I insist one needn't be conscious of his or her reasons). But it's not obvious such esoteric science connects up with anything you would have recognized as a reason, prior to substantial educational intervention. This suggests that the counterfactual test needs to be amended, with a restriction that the counterfactual recognition not require too substantial a change to your cognitive and motivational structures at the time of action. Once again, I'm pressed to think harder about what cognitive processes deserve to be called "reasoning."

Finally, my approach bears strong affinities to Alfano's (2013a) suggestive account of moral character, where, as stated in **Alfano**'s commentary here, "tactically deployed fictions about ourselves can become facts" (e.g., falsely attributing honesty to oneself can help one behave more honestly in the future).[12] As with my collaborativism, Alfano thinks this process is substantially social, involving iterated "bid-and-accept patterns" where a person announces what "her values, motives, concerns, or drives are" for evaluation and ratification (or, presumably, rejection) by her associates. The process may also go in the other direction: a person may announce what another person is, and that other may accept or reject the attribution.

From here, **Alfano** makes the intriguing suggestion that we must recognize "a novel class of dispositions – namely, the dispositions associated with being a good echo" – a person who helps others live up to their announced self. If these dispositions are admirable dispositions, they must also involve a readiness not to echo but oppose – for instance, when one's interlocutor makes pernicious announcements, such as those that are harmful to self and others (here, as elsewhere, the agential and the morally good may come apart). Furthermore, as Alfano knows better than most, whatever theory one wishes to construct for these echo-dispositions, it must account for the fact that dispositions are highly liable to situational disruptions. Nevertheless, Alfano is right to think that as we attempt to more completely uncover the psychological mechanisms supporting collaborativism, we must look for the relevant individual dispositional differences, even if they are not so robustly impactful as we might have hoped.

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

R3. What are theories of agency for?

R3.1. Agency and the law

Mattei contends that if the position I develop "is to be taken seriously . . . the proposed conceptual framework should be able to transcend the purely theoretical realm." It's not obvious that all theorizing must have practical implications to be "taken seriously" (such a demand seems singularly inapposite for work in metaphysics, for example), but I accept this aspiration for my own work, which I expect might inform, and be informed by, the everyday practice of responsibility attribution (pp. 5, 156–59). From there, Mattei and I part ways, because he considers only one practical domain, that of law, while my project is officially neutral on topics like criminal responsibility (p. 24). My focus throughout was on how to think about everyday interpersonal relationships, and because it seems quite undeniable such relationships are of great practical significance, I suppose that the theorizing I do meets such standards of practical relevance as are appropriate for theorizing in moral psychology and ethics.

I do not know if my theory is, as **Mattei** says, "irreconcilable with key principles of the American common law tradition." To re-emphasize, this issue is avowedly not my issue, but I'll say a little about it. Assuming Mattei's charge of irreconcilability is right, we've a couple of interesting possibilities: so much the worse for my theory, or so much the worse for the law. Now the relationship between moral responsibility and criminal responsibility is a delicate one; but while matters are controversial, it seems at least to be commonly held that criminal responsibility should "track" moral responsibility (Duff 2009; problematic issues like strict liability noted). If this is right, I can say that the law should be accountable to the best going theory of moral responsibility rather than the other way around, and it's unclear why any tension should be mine to ameliorate.

Suppose one asserted the contrary, as we might take **Mattei** to do. I'd want to proceed by considering concrete cases, but I'd expect to find many places where we'd want to resist the law dictating our theory of agency. For example, given psychopaths' profound deficits in emotional processing and impulse control (Kiehl 2014), it is highly plausible that they are not morally responsible for what they do. Yet the law does not excuse or exempt psychopaths (who may account for 25% of the prison population), and it is convincingly argued that this state of affairs is unjust (Morse 2008). Whether or not you agree – maybe the extreme dangerousness of psychopaths justifies their incarceration – there seems scant ground for letting the current state of criminal law determine our thinking on the moral responsibility of psychopaths, given what is known about the illness.

Perhaps there are instances where my theory of moral responsibility would apply infelicitously to questions of legal responsibility, as **Mattei** insists. But as a pluralist (pp. 171–77), this doesn't trouble me; I deny that any single theory of responsibility is equally applicable across all contexts, and I'd not be surprised if notions of responsibility that do good work in many everyday contexts, like interpersonal relationships, are not applicable to the particular context of the law. In any event, there is much to be learned by considering moral responsibility independently of criminal responsibility, as a great many philosophers have done (even if it can be shown that the domains importantly interact).

3.2. In Praise of Busybodies

Sommers fears not that my pragmatic aspirations are too limited, but that they are too extensive. He proposes that I should, instead of adopting a pluralism that accommodates numerous theoretical perspectives, adopt a pluralism that "rejects theorizing about responsibility altogether." According to Sommers, it's not "the philosopher's business to cast judgment" on everyday judgments of responsibility; to do so is to be a philosophical "busybody," meddling in practices that successfully deliver what participants in the practices require.[13]

Philosophers, at least since Socrates, the patron saint of philosophical busybodies, have staged "philosophical interventions" on theoretically suspect practices and beliefs. So **Sommers**'s position is at once conservative, because it argues that responsibility theorists should not attempt to alter existing practice, and radical, because it advocates overturning what is plausibly thought to be the animating spirit of the philosophical enterprise.

In defense of his view, **Sommers** offers Lee, from the film *Manchester by the Sea*, whose drunken carelessness causes his children to perish in a fire. According to Sommers, Lee would not be held responsible on my valuational theory, because the deaths of his children clearly do not express his values. I could dispute this: maybe Lee's behavior ought be understood as stemming from his valuing partying too much, and the responsibilities of parenting too little, and maybe then, valuational theory has him responsible.

But the main question for **Sommers** is not Lee's responsibility, but who gets to decide. Sommers thinks responsibility attributions are for "the people who are involved in the situation to arrive at," while I think the theorist has a say. However, I'm not arguing that the role of the theorist is to engage in direct intervention. Sommers is probably right that abstract philosophical theorizing is unlikely to help Lee with his crushing guilt (and/or shame), which we may suppose reflects a self-attribution of responsibility. For that kind of intervention, another kind of busybody, the therapist, is likely better suited (assuming ameliorating Lee's pain would be a good outcome, which Sommers may deny).

Suppose I'm right that many forms of mental illness can be understood as impairments of agency (pp. 34–35). Now allow me to speculate that many sufferers of mental illness are held morally responsible for illness-related behaviors, either because they are not recognized as ill, or because the connection between their illness and impairments of agency are insufficiently appreciated. If this is right, many responsibility attributions of "people who are involved in the situation," are likely to be quite wrong by the standards of the best going theory. Still, the theorist shouldn't expect to directly change minds, or change them overnight. Rather, any practical influence of theory is likely to be diffuse, indirect, and long-term. Theoretical influence might proceed, for example, through popular forms of media; lots of people flip through *Psychology Today* at the grocery store checkout, and the magazine apparently boasts a circulation of 275,000 (probably not too shabby,

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

in the internet era). Moreover, the practical influence of theory may be *salutary* influence; if academic theory has had a role in more humane attitudes toward mental illness, that's a good thing. We ought, I think, be thankful for philosophical busybodies, even if we have good reason to ignore some of their recommendations.[14]

NOTES

1. I explain my use of "agency" in sect. 1 of the précis.

2. Unless otherwise noted, all references to my work are for *Talking to Our Selves*, and by page number only, and all references to my commentators are for the contributions to this symposium.

3. "Small" gets used to different purposes in this vicinity, and we should distinguish effect size and probability: small effect sizes are not less probable than large effect sizes.

4. **Levy** apparently agrees that my skeptical problem *does* trouble reflectivist theories; it is therefore unclear, given the philosophical prominence of reflectivism (pp. 17–19), why he should deny the problem's importance.

5. **Ward & Machery**'s "only require," seems to intimate sufficiency; if it does not, we require an account of what else is required.

6. For example, see Harman (1965, p. 89) on the better explanation being the "more plausible" explanation. I used to think this fudging. I'm now of the view that such appeals to "horse sense" are ineliminable; the best one can do is situate them in a compelling theory. Different theorists (and different cultures) will of course differ on where such expedients are acceptable.

7. In an instructive paper, Keijzer (2013) argues that the philosophers have overstated the *Sphex's* roboticism. But even on Keijzer's reassessment, the wasps seem pretty darn mechanical.

8. **Fowers et al.** contend that I "failed to accurately describe" the "individual differences result," in Cameron et al. (2013). I cited this paper (p. 54), together with other work, in making the well-substantiated observation that emotions can influence moral judgment. Fowers et al. note there was no main effect for disgust in Cameron et al. (2013), but also that some participants (those low in emotion differentiation) were affected by disgust. That is, my reference to Cameron et al. (2013) was accurate.

9. On the other hand, many people apparently assume, as **Niemi & Graham** note, that others are to be protected from "painful imposition," so perhaps the negotiation of responsibility attributions may be resolved by appeal to this common ground.

10. Following are a few examples of this discussion, sampled more or less randomly from the internet:
https://www.theguardian.com/commentisfree/2016/dec/06/louise-casey-discrimination-muslim-women-bradford
http://www.annemariewaters.org/sharia-law-and-middle-class-feminism/
https://briarpatchmagazine.com/articles/view/racism-feminism-and-the-sharia-debate
I do not claim expertise on this issue, nor do I forward any opinion on it other than to note that the debate may help illustrate the delicacy of thinking about adaptive preferences and agency.

11. Previously, my own readings trended to the provocative: Doris (2002, pp. 105–106; Doris 2005, pp. 673–74).

12. I should also acknowledge **Alfano**'s observation that my account of agency bears considerable affinities to Nietzsche's.

13. This "quietism" about everyday practice may place **Sommers**'s views closer to Strawson (1962) than mine are, since I reject Strawson's (apparent) quietism.

14. Many thanks to Paul Bloom, Casey O'Callaghan, and Lauren Olin for valuable comments on an earlier version of this response. Much of the writing was done during a term as a Laurence S. Rockefeller Fellow at Princeton's University Center for Human Values. I'm most grateful to the Center, and to Washington University in St. Louis for sabbatical leave.

References

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

Ahrens, M. (2016) *Home fires involving cooking equipment*. NFPA Fire Analysis and Research. [SM]

Alesina, A. & Angeletos, G.-M. (2005) Fairness and redistribution. *The American Economic Review* 95(4):960–80. [HV]

Alex, K. (2008) Vicarious criminal liability and the constitutional dimensions of Pinkerton. *American University Law Review* 57:585–639. [TAM]

Alexander, R. D. (1987) *The biology of moral systems*. Aldine de Gruyter. [PMP]

Alfano, M. (2010) The tenacity of the intentional prior to the *Genealogy*. *Journal of Nietzsche Studies* 40:123–40. [MA]

Alfano, M. (2013a) *Character as moral fiction*. Cambridge University Press. [MA, rJMD, AP-C]

Alfano, M. (2013b) Nietzsche, naturalism, and the tenacity of the intentional. *Journal of Nietzsche Studies* 44(3):457–64. [MA]

Alfano, M. (2015) How one becomes what one is called: On the relation between traits and trait-terms in Nietzsche. *Journal of Nietzsche Studies* 46(1):261–69. [MA]

Alfano, M. (2016a) Friendship and the structure of trust. In: *From personality to virtue*, ed. A. Masala & J. Webber, pp. 186–206. Oxford University Press. [MA]

Alfano, M. (2016b) How one becomes what one is: The case for a Nietzschean conception of character development. In: *Questions of character*, ed. I. Fileva, 89–104. Oxford University Press. [MA]

Alfano, M. (2016c) *Moral psychology: An introduction*. Polity. [MA]

Alicke, M. D. (2000) Culpable control and the psychology of blame. *Psychological Bulletin* 126(4):556–74. doi:10.1037/0033-2909.126.4.556. [PMP]

Alicke, M. D., Vredenburg, D. S., Hiatt, M. & Govorun, O. (2001) The "better than myself" effect. *Motivation and Emotion* 25(1):7–22. [aJMD]

Amaya, S. (2013) Slips. *Noûs* 47(3):559–76. [SM]

Amaya, S. (2015) The argument from slips. In: *Agency, freedom, and moral responsibility*, ed. A. Buckareff, C. Moya & S. Rosell, pp. 13–29. Palgrave Macmillan. [SM]

Amaya, S. (2016) Slip-proof actions. In: *Time and the philosophy of action*, ed. R. Altshuler & M. J. Sigrist, pp. 21–36. Routledge. [SM]

Amaya, S. & Doris, J. M. (2015) No excuses: Performance mistakes in morality. In: *Handbook of neuroethics*, ed. J. Clausen & N. Levy, pp. 352–71. Springer. [rJMD]

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C.R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., van der Hulst, M., Jonas, K. J., Lai, C. K., Levitan, C. A., Miller, J. K., Moore, K. S., Meixner, J. M., Munafò, M. R., Neijenhuijs, K. I., Nilsonne, G., Nosek, B. A., Plessow, F., Prenoveau, J. M., Ricker, A. A., Schmidt, K., Spies, J. R., Stieger, S., Strohminger, N., Sullivan, G. B., van Aert, R. C., van Assen, M. A., Vanpaemel, W., Vianello, M., Voracek, M. & Zuni, K. (2016) Response to comment on "estimating the reproducibility of psychological science." *Science* 351(6277):1037. [aJMD]

Andreoni, J., Rao, J. M. & Trachtman, H. (in press) Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*. NBER Working paper No. 17648, issued December 2011, revised November 2016. DOI: 10.3386/w17648. [DD]

Annas, J. (1993) *The morality of happiness*. Oxford University Press. [DD]

Anscombe, G. E. M. (1957) *Intention*. Harvard University Press. [ZBW]

Appiah, K. A. (2010) *The ethics of identity*. Princeton University Press. [aJMD]

Ariely, D. (2012) *The honest truth about dishonesty*. Harper Collins. [JM]

Aristotle (340 BCE/1999). *Nicomachean ethics*, trans. M. Ostwald. Upper Saddle River, NJ: Prentice Hall. (Original work published around 340 BCE.) [BJF]

Arpaly, N. (2002) *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press. [aJMD]

Arpaly, N. (2003) *Unprincipled virtue*. Oxford University Press. [JM]

Arpaly, N. (forthcoming) Comments on *Talking to Our Selves* by John Doris. *Philosophy and Phenomenological Research*. [rJMD]

Arpaly, N. & Schroeder, T. (1999) Praise, blame, and the whole self. *Philosophical Studies* 93(2):161–88. [SB]

Arpaly, N. & Schroeder, T. (2012) Deliberation and acting for reasons. *Philosophical Review* 121(2):209–39. [ZBW]

Ayduk, O., Mendoza-Denton, R., Mischel, W., Downey, G., Peake, P. K. & Rodriguez, M. (2000). Regulating the interpersonal self: Strategic self-regulation for coping with rejection sensitivity. *Journal of Personality and Social Psychology* 79:776–92. [BJF]

Back, M. D., Schmukle, S. C. & Egloff, B. (2009) Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology* 97(3):533. [ZBW]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Baker, L. A. & Emery, R. E. (1993) When every relationship is above average: Perceptions and expectations of divorce at the time of marriage. *Law and Human Behavior* 17(4):439. [aJMD]

Balliet, D., Mulder, L. B. & Van Lange, P. A. M. (2011) Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin* 137(4):594–615. doi:10.1037/a0023489. [SH]

Bandura, A. (1990) Selective activation and disengagement of moral control. *Journal of Social Issues* 46(1):27–46. doi:10.1111/j.1540-4560.1990.tb00270.x. [PMP]

Bargh, J. A. (1994) The Four Horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In: *Handbook of social cognition*, 2nd edition, ed. R. S. Wyer, Jr. & T. K. Srull, pp. 1–40. Erlbaum. [JSU]

Bargh, J. A. & Chartrand, T. L. (1999) The unbearable automaticity of being. *American Psychologist* 54:462–79. [MT]

Baron, R. (2005) So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. *Advances in Experimental Social Psychology* 37:219–53. doi: 10.1016/S0065-2601(05)37004-3. [EL]

Baron, J. & Spranca, M. (1997) Protected values. *Organizational Behavior and Human Decision Processes* 70:1–16. doi:10.1006/obhd.1997.2690. [SH]

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisorf, A., Scelzaa, B.A., Stichl, S., von Ruedenn, C., Zhaoh, W. & Laurence, S. (2016) Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences* 113(17):4688–93. Available at: https://doi.org/10.1073/pnas.1522070113. [HV]

Bateson, M., Nettle, D. & Roberts, G. (2006) Cues of being watched enhance cooperation in a real-world setting. *Biology Letters* 2(3):412–14. [JM]

Batson, C. D. (2011) *Altruism in humans*. Oxford University Press. [JM]

Baumeister, R. F., Bratslavsky, E., Muraven, M. & Tice, D. M. (1998) Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74:1252–65. [MT]

Baumeister, R. F., Masicampo, E. J. & DeWall, C. N. (2009) Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin* 35(2):260–68. doi:10.1177/0146167208327217. [PMP]

Beevers, C. G. (2005) Cognitive vulnerability to depression: A dual process model. *Clinical Psychology Review* 25(7):975–1002. [aJMD]

Benabou, R. & Tirole, J. (2006) Belief in a just world and redistributive politics. *The Quarterly Journal of Economics* 121(2):699–746. [HV]

Berlyne, N. (1972) Confabulation. *British Journal of Psychiatry* 120:31–39. [DD]

Berthold, A., Mummendey, A., Kessler, T., Luecke, B. & Schubert, T. (2012) When different means bad or merely worse. How minimal and maximal goals affect ingroup projection and outgroup attitudes. *European Journal of Social Psychology* 42:682–90. doi:10.1002/ejsp.1878. [SH]

Bickhard, M. (2016) Inter- and En-activism: Some thoughts and comparisons. *New Ideas in Psychology* 41:23–32. [SIH]

Bigelow, A. E. & Rochat, P. (2006) Two-month-old infants' sensitivity to social contingency in mother–infant and stranger–infant interaction. *Infancy* 9(3):313–25. [BB]

Blackburn, R. T. & Clark, M. J. (1975) An assessment of faculty performance: Some correlates between administrator, colleague, student and self-ratings. *Sociology of Education* 48(2):242–56. [aJMD]

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Gusman, F. D., Charney, D. S. & Keane, T. M. (1995) The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress* 8:75–90. [LN]

Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., Kleutsch, L., Kramer, K. L., Ross, E., Vongsachang, H., Wrangham, R. & Warneken, F. (2015) The ontogeny of fairness in seven societies. *Nature* 528 (7581):258–61. [BB]

Blasi, A. (2005) Moral character: A psychological approach. In: *Character psychology and character education*, ed. D. K. Lapsley & F. C. Clark, pp. 67–100. Notre Dame University Press. [BJF]

Bloom, P. (2014) The war on reason. *The Atlantic*. Available at: https://www.theatlantic.com/magazine/archive/2014/03/the-war-on-reason/357561/ [aJMD]

Bloom, P. (2016) *Against empathy.* Bodley Head. [BF]

Bollich, K. L., Doris, J. M., Vazire, S., Raison, C. L., Jackson, J. J. & Mehl, M. R. (2016) Eavesdropping on character: Testing the stability of naturalistically observed daily moral behavior. *Journal of Research in Personality* 61:15–21. [rJMD]

Boyer-Pennington, M. E., Pennington, J. & Spink, C. (2001) Students' expectations and optimism toward marriage as a function of parental divorce. *Journal of Divorce & Remarriage* 34(3–4):71–87. [aJMD]

Bratman, M. (2000) Valuing and the will. *Philosophical Perspectives* 14:249–65. [SM]

Bratman, M. E. (2007) *Structures of agency: Essays*. Oxford University Press. [SB, aJMD, MRV]

Bray, J. H. & Jouriles, E. N. (1995) Treatment of marital conflict and prevention of divorce. *Journal of Marital and Family Therapy* 21(4):461–73. [aJMD]

Brenner, B. (1979) Depressed affect as a cause of associated somatic problems. *Psychological Medicine* 9(4):737–46. [aJMD]

Brewer, M. B. & Gardner, W. L. (1996) Who is this "We"? Levels of collective identity and self representations. *Journal of Personality and Social Psychology* 71(1):83–93. [BF]

Brink, D. O. (1992) Mill's deliberative utilitarianism. *Philosophy and Public Affairs* 21:67–103. [DD]

Brink, D. O. & Nelkin, D. (2013) Fairness and the architecture of responsibility. *Oxford Studies in Agency and Responsibility* 1:284–313. [rJMD]

Brown v. Kendall (1850) 60 Mass. (6 Cush.) 292. Available at: http://moglen.law.columbia.edu/twiki/pub/EngLegalHist/MitchellAllestry/Brown_v_Kendall.pdf. [TAM]

Brown, J., Dreis, S. & Nace, D. K. (1999) What really makes a difference in psychotherapy outcome? Why does managed care want to know? In: *The heart and soul of change: What works in therapy*, ed. M. A. Hubble, B. L. Duncan & S. D. Miller, pp. 389–406. American Psychological Association. [aJMD]

Brown, P. & Levinson, S. C. (1987) *Politeness: Some universals in language usage*. Cambridge University Press. [DD, JZ]

Brueckner, A. (1994) The structure of the skeptical argument. *Philosophy and Phenomenological Research* 54(4):827–35. [JM]

Bushman, B. J., DeWall, C. N., Pond, R. S. & Hanus, M. D. (2014) Low glucose relates to greater aggression in married couples. *Proceedings of the National Academy of Sciences* 111(17):6254–57. [ZBW]

Cain, D. M., Dana, J. & Newman, G. E. (2014) Giving versus giving in. *The Academy of Management Annals* 8(1):505–33. [DD]

Calhoun, C. (2004) An apology for moral shame. *Journal of Political Philosophy* 12(2):127–46. [AP-C]

Cameron, C. D., Payne, B. K. & Doris, J. M. (2013) Morality in high definition: Emotion differentiation calibrates the influence of incidental disgust on moral judgments. *Journal of Experimental Social Psychology* 49(4):719–25. Available at: https://doi.org/10.1016/j.jesp.2013.02.014. [rJMD, BJF]

Campbell, R. L. & Bickhard, M. H. (1986) *Knowing levels and developmental stages*. Karger. [SIH]

Carlson, M., Charlin, V. & Miller, N. (1988) Positive mood and helping behavior: A test of six hypotheses. *Journal of Personality and Social Psychology* 55(2):211–29. [JM]

Carpendale, J. I. M. (2000) Kohlberg and Piaget on stages and moral reasoning. *Developmental Review* 20:181–205. [SIH]

Carpendale, J. I. M. (2009) Piaget's theory of moral development. In: *The Cambridge companion to Piaget*, ed. U. Müller, J. I. M. Carpendale & L. Smith, pp. 270–86. Cambridge University Press. [SIH]

Carpendale, J. I. M., Hammond, S. I. & Atwood, S. (2013) A relational developmental systems approach to moral development. In: *Advances in child development and behavior, vol. 44,* ed. R. M. Lerner & J. B. Benson, pp. 125–53. Elsevier Science. [SIH]

Carruthers, P. (2009) How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32:121–82. [JJC]

Carson, C. D & Felthous, A. R. (2003) Mens rea. *Behavioral Sciences and the Law* 21:559–62. [TAM]

Caruso, G. D. (2014) (Un)just deserts: The dark side of moral responsibility. *Southwest Philosophy Review* 30(1):27–38. doi:10.5840/swphilreview20143014. [PMP]

Cary, P. (2007) A brief history of the concept of free will: Issues that are and are not germane to legal reasoning. *Behavioral Sciences and the Law* 25:165–81. [TAM]

Catholic Church (2000) *Catechism of the Catholic Church popular revised edition*. Continuum. [DC]

Chandler, M J., Lalonde, C E., Sokol, B W. & Hallett, D. (2003) Personal persistence, identity development, and suicide: A study of native and non-native North American adolescents. *Monographs of the Society for Research in Child Development* 68:1–130. [SIH]

Chapman, M. (1988) *Constructive evolution: Origins and development of Piaget's thought*. Cambridge University Press. [SIH]

Cherlin, A. J., Furstenberg, F. F., Chase-Lansdale, L., Kiernan, K. E., Robins, P. K., Morrison, D. R. & Teitler, J. O. (1991) Longitudinal studies of effects of divorce on children in Great Britain and the United States. *Science* 252(5011):1386–89. [ZBW]

Chi, M. T. (2006) Two approaches to the study of experts' characteristics. In: *The Cambridge handbook of expertise and expert performance*, ed. K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman, pp. 21–30. Cambridge University Press. [rJMD]

Choudhry, N. K., Fletcher, R. H. & Soumerai, S. B. (2005) Systematic review: The relationship between clinical experience and quality of health care. *Annals of Internal Medicine* 142:260–73. [MT]

Christensen, W., Sutton, J. & McIlwain, D. J. (2016) Cognition in skilled action: Meshed control and the varieties of skill experience. *Mind & Language* 31(1):37–66. [arJMD]

Cialdini, R. (2008) *Influence: Science and practice*, 5th edition. Allyn and Bacon. [EL]

Cialdini, R. B. & Goldstein, N. J. (2004) Social influence: Compliance and conformity. *Annual Review of Psychology* 55:591–621. doi:10.1146/annurev.psych.55.090902.142015. [SH]

Cialdini, R. B., Vincent, J. E., Lewis, S. K., Catalan, J., Wheeler, D. & Darby, B. L. (1975) Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of Personality and Social Psychology* 31:206–15. [EL]

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F. & Baumeister, R. F. (2014) Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology* 106(4):501–13. doi:10.1037/a0035880. [PMP]

Clifford, S., Jewell, R. M. & Waggoner, P. D. (2015) Are samples drawn from Mechanical Turk valid for research on political ideology? *Research and Politics* 2 (4). Available at: https://doi.org/10.1177/2053168015622072. [LN]

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*, 2nd edition. Erlbaum. [arJMD]

Cohen, L. B. & Salapatek, P., eds. (2013) *Infant perception: From sensation to cognition: Basic visual processes, vols. 1 & 2.* Academic Press. [DC]

Collerton, D., Perry, E. & McKeith, I. (2005) Why people see things that are not there: a novel perception and attention deficit model for recurrent complex visual hallucinations. *Behavioral and Brain Sciences* 28(6):737–57. [DC]

Collins, J. (2006) Crime and parenthood: The uneasy case of prosecuting negligent parents. *Northwestern University Law Review* 100:807–56. [rJMD]

Couchman, J. J. (2012) Self-agency in rhesus monkeys. *Biology Letters* 8(1):39–41. doi: 10.1098/rsbl.2011.0536. [JJC]

Couchman, J. J., Coutinho, M. V. C., Beran, M. J. & Smith, J. D. (2009) Metacognition is prior. *Behavioral and Brain Sciences* 32(2):142. [JJC]

Couchman, J. J., Miller, N., Zmuda, S. J., Feather, K. & Schwartzmeyer, T. (2016) The instinct fallacy: The metacognition of answering and revising during college exams. *Metacognition & Learning* 11(2):171–85. [JJC]

Coutinho, M. V. C., Redford, J. S., Church, B. A., Zakrzewski, A. C., Couchman, J. J. & Smith, J. D. (2015) The interplay between uncertainty monitoring and working memory: Can metacognition become automatic? *Memory & Cognition* 43(7):990–1006. [JJC]

Cramer, R. E., McMaster, M. R., Bartell, P. A. & Dragna, M. (1988) Subject competence and minimization of the bystander effect. *Journal of Applied Social Psychology* 18:1133–48. [rJMD, MT]

Crockett, M. J., Clark, L., Hauser, M. D. & Robbins, T. W. (2010) Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *PNAS* 107(40):17433–38. [LN]

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., Grosse-Rueskamp, J. M., Dayan, P. & Dolan, R. J. (2015) Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. *Current Biology* 25(14):1852–59. [DC]

Cross, K. P. (1977) Not can, but will college teaching be improved? *New Directions for Higher Education* 17:1–15. [aJMD]

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

Cushman, F. & Greene, J. D. (2012) The philosopher in the theater. In: *The social psychology of morality: exploring the causes of good and evil*, ed. M. Mikulincer & P. R. Shaver, pp. 33–50. APA Press. [PMP]

Cushman, F., Young, L. & Greene, J. D. (2010) Multi-system moral psychology. In: *The moral psychology handbook*, ed. J. M. Doris & the Moral Psychology Research Group, pp. 47–71. Oxford University Press. [aJMD]

D'Arms, J. & Jacobson, D. (2000) The moralistic fallacy: On the "appropriateness" of emotions. *Philosophical and Phenomenological Research* 61(1):65–90. [aJMD, HV]

D'Arms, J. & Jacobson, D. (2006) Anthropocentric constraints on human value. *Oxford Studies in Metaethics* 1:99–126. [aJMD]

Dale, R., Fusaroli, R., Duran, N. D. & Richardson, D. C. (2013) The self-organization of human interaction. In: *Psychology of learning and motivation*, ed. B. Ross, pp. 43–95. Academic Press. [SB]

Daly, M. & Wilson, M. (1996) Violence against stepchildren. *Current Directions in Psychological Science* 5:77–81. [rJMD]

Daly, M. & Wilson, M. (2007) Is the "Cinderella effect" controversial? A case study of evolution-minded research and critiques thereof. In: *Foundations of evolutionary psychology*, ed. C. Crawford & D. Krebs, pp. 383–400. Erlbaum. [rJMD]

Darley, J. M., Carlsmith, K. M. & Robinson, P. H. (2000) Incapacitation and just deserts as motives for punishment. *Law and Human Behavior* 24(6):659–83. doi:10.1023/A:1005552203727. [SH]

Darley, J. M. & Latané, B. (1968) Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* 8:(4, Pt. 1):377–83. [aJMD, MT]

Davies, P. S. (2009) *Subjects of the world: Darwin's rhetoric and the study of agency in nature*. University of Chicago Press. [aJMD]

Davis, D., Sundahl, I. & Lesbo, M. (2000) Illusory personal control as a determinant of bet size and type in casino craps games. *Journal of Applied Social Psychology* 30(6):1224–42. [aJMD]

Dawes, R. M. (1994) Psychotherapy: The myth of expertise. In: *House of cards: Psychology and psychotherapy built on myth*, ed. R. M. Dawes, pp. 38–74. Free Press. [aJMD]

Dawson, E., Gilovich, T. & Regan, D. T. (2002) Motivated reasoning and performance on the Wason card election task. *Personality and Social Psychology Bulletin* 28(10):1379–87. [aJMD]

Debove, S., Baumard, N. & André, J.-B. (2015) Evolution of equal division among unequal partners. *Evolution* 69(2):561–69. Available at: https://doi.org/10.1111/evo.12583. [HV]

Deffains, B., Espinosa, R. & Thöni, C. (2016) Political self-serving bias and redistribution. *Journal of Public Economics* 134:67–74. doi:10.1016/j.jpubeco.2016.01.002. [PMP]

Demaree-Cotton, J. (2016) Do framing effects make moral intuitions unreliable? *Philosophical Psychology* 29(1):1–22. [JM]

Dennett, D. C. (1984) *Elbow room: The varieties of free will wanting*. Oxford University Press. [rJMD, PMP]

Dennett, D. C. (1991) The reality of selves. In: *Consciousness explained*, ed. D. C. Dennett, pp. 412–30. Little, Brown. [aJMD]

Dennett, D. C. (1992) The self as a center of narrative gravity. In: *Self and consciousness: Multiple perspectives*, ed. F. S. Kessel, P. M. Cole & D. L. Johnson, pp. 103–15. Erlbaum. [aJMD]

Descartes, R. (1641/2008) *Meditations on first philosophy: With selections from the objections and replies*. Oxford University Press. (Original work published in 1641.) [rJMD]

DeWall, C. N., Baumeister, R. F., Gailliot, M. T. & Maner, J. K. (2008) Depletion makes the heart grow less helpful: Helping as a function of self-regulatory energy and genetic relatedness. *Personality and Social Psychology Bulletin* 34:1663–76. [MT]

Diener, E. & Chan, M. Y. (2011) Happy people live longer: Subjective well-being contributes to health and longevity. *Applied Psychology: Health and Well-Being* 3(1):1–43. [aJMD]

Ditto, P. H. & Lopez, D. F. (1992) Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology* 63(4):568–84. [aJMD]

Divorce Magazine. (2004) U.S. divorce statistics [data file]. Retrieved from DivorceMagazine.com: http://www.divorcemag.com/statistics/statsUS.shtml. [aJMD]

Dixon, R. & Gould, O. (1996) Adults telling and retelling stories collaboratively. In: *Interactive minds: Life-span perspectives on the social foundation of cognition*, ed. P. Baltes & U. Staudinger, pp. 221–41. Cambridge University Press. [MA]

Doris, J. M. (2002) *Lack of character: Personality and moral behavior*. Cambridge University Press. [BB, arJMD, BJF, SIH, HLM, AP-C]

Doris, J. M. (2005) Replies: Evidence and sensibility. *Philosophy and Phenomenological Research* 72:656–77. [rJMD]

Doris, J. M. (2009) Skepticism about persons. *Philosophical Issues* 19(1):57–91. [AP-C]

Doris, J. M. (2015a) Doing without (arguing about) desert. *Philosophical Studies* 172 (10):2625–34. [aJMD]

Doris, J. M. (2015b). *Talking to our selves: Reflection, ignorance, and agency*. Oxford University Press. [BB, SB, DC, arJMD, BJF, SIH, WH, EL, NL, HLM, JM, SM, TAM, LN, PMP, AP-C, TS, MT, JSU, HV, MRV, ZBW, JZ]

Doris, J. M. (forthcoming). *Character trouble: Undisciplined essays on agency and personality*. Oxford University Press. [rJMD]

Doris, J. M. (in preparation) Making good: In search of moral expertise. [rJMD]

Doris, J. M. & Murphy, D. (2007) From My Lai to Abu Ghraib: The moral psychology of atrocity. *Midwest Studies in Philosophy* 31(1):25–55. [rJMD]

Doris, J. M. & Plakias, A. (2007) How to argue about disagreement: Evaluative diversity and moral realism. In: *Moral psychology, vol. 2, The cognitive science of morality*, ed. W. Sinnott-Armstrong, pp. 303–31. MIT Press. [rJMD]

Dreyfus, H. L. (1985) Holism and hermeneutics. In: *Hermeneutics and praxis*, ed. R. Hollinger, pp. 227–47. University of Notre Dame Press. [BF]

Duff, A. (2009) Legal and moral responsibility. *Philosophy Compass* 4(6):978–86. [rJMD]

Duff, R. A. (1990) *Intention, agency and criminal liability: Philosophy of action and the criminal*. Blackwell. [TAM]

Dufner, M., Denissen, J. J., Zalk, M., Matthes, B., Meeus, W. H., van Aken, M. A. & Sedikides, C. (2012) Positive intelligence illusions: On the relation between intellectual self-enhancement and psychological adjustment. *Journal of Personality* 80(3):537–72. [aJMD]

Dunning, D. (1999) A newer look: Motivated social cognition and the schematic representation of social concepts. *Psychological Inquiry* 10(1):1–11. [aJMD]

Dunning, D. (2006) *Self-insight: Roadblocks and detours on the path to knowing thyself*. Psychology Press. [aJMD]

Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D. & Fetchenhauer, D. (2014) Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology* 107:122–41. [DD]

Dunning, D. & Fetchenhauer, D. (2013) Behavioral influences in the present tense: On expressive versus instrumental action. *Perspectives on Psychological Science* 8:142–45. [DD]

Dunning, D., Fetchenhauer, D. & Schlösser, T. (2016) The psychology of respect: A case study of how behavioral norms regulate human action. In: *Advances in motivation science, vol. 3*, ed. A. Elliot, pp. 1–34. Elsevier. [DD]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Dunning, D., Heath, C. & Suls, J. (2004) Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest* 5:71–106. [DD]

Dunning, D., Kunda, Z. & Murray, S. L. (1999) What the commentators motivated us to think about. *Psychological Inquiry* 10(1):79–82. [aJMD]

Durkheim, É. (1893/1996) *Über soziale Arbeitsteilung: Studie über die Organisation höherer Gesellschaften,* 2nd edition. Suhrkamp. (Original work published in 1893.) [SH]

Ebbinghaus, H. (1913) *On memory: A contribution to experimental psychology*. Teachers College. [JJC]

Ehlers, A. & Clark, D. M. (2000) A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy* 38:319–45. [LN]

Elster, J. (1983) *Sour grapes: Studies in the subversion of rationality*. Cambridge University Press. [MRV]

Ericsson, K. A. (2014) Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms. *Intelligence* 45:81–103. [rJMD]

Ericsson, K. A., Whyte, J. & Ward, P. (2007) Expert performance in nursing: Reviewing research on expertise in nursing within the framework of the expert-performance approach. *ANS Advances in Nursing Science* 30:E58–E71. [MT]

Evans, J. S. B. & Frankish, K. E. (2009) *In two minds: Dual processes and beyond*. Oxford University Press. [aJMD]

Evans, J. S. B. & Over, D. E. (1996) *Rationality and reasoning*. Psychology Press. [aJMD]

Farrell, C., Cowley, E. & Edwardson, M. (2005) Strategies to improve the probability of winning a lottery: Gamblers and their illusions of control. *European Advances in Consumer Research* 7:597–601. [aJMD]

Feinberg, J. (1965) The expressive function of punishment. *The Monist* 49:397–423. doi:10.5840/monist196549326. [SH]

Feldman, G., Chandrashekar, S. P. & Wong, K. F. E. (2016) The freedom to excel: Belief in free will predicts better academic performance. *Personality and Individual Differences* 90:377–83. doi:10.1016/j.paid.2015.11.043. [PMP]

Fessler, D. M. & Holbrook, C. (2013) Baumard et al.'s moral markets lack market dynamics. *Behavioral and Brain Sciences* 36(1):89–90. [HV]

Fiore, M. C., Bailey, W. C., Cohen, S. J., Dorfman, S. F., Goldstein, M. G., Gritz, E. R., Heyman, R. B., Jaen, C. R., Kottke, T. E., Lando, H. A. & Mecklenburg, R. E. (2000) *Treating tobacco use and dependence: Clinical practice guideline*. U.S. Department of Health and Human Services. [aJMD]

Fischer, J. M. (2006) *My way: Essays on moral responsibility*. Oxford University Press. [aJMD]

Fischer, J. M. (2009) *Our stories: Essays on life, death, and free will*. Oxford University Press. [aJMD]

Fischer, J. M. (forthcoming). On John Doris's *Talking to Our Selves. Social Theory and Practice*. [rJMD]

Fischer, J. M. & Ravizza, M. (1998) *Responsibility and control: A theory of moral responsibility*. Cambridge University Press. [aJMD, NL, SM, MRV]

Fischhoff, B. (2007) An early history of hindsight research. *Social Cognition* 25:10–13. [JJC]

Fiske, A. P. & Rai, T. S. (2014) *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press. [HV]

Floyd, S., Rossi, G., Enfield, N. J., Baranova, J., Blythe, J., Dingemanse, M., Kendrick, K. & Zinken, J. (2014). Recruitments across languages: A systematic comparison. Talk presented at the *4th International Conference on Conversation Analysis (ICCA 2014). University of California at Los Angeles, CA, June 25–29, 2014.* [JZ]

Foa, E. B., Ehlers, A., Clark, D. M., Tolin, D. F. & Orsillo, S. M. (1999) The posttraumatic cognitions inventory (PTCI): Development and validation. *Psychological Assessment* 11(3):303–14. [LN]

Forgas, J. P., Dunn, E. & Granland, S. (2008) Are you being served…? An unobtrusive experiment of affective influences on helping in a department store. *European Journal of Social Psychology* 38:333–42. [MT]

Förster, J., Liberman, N. & Friedman, R. S. (2009) What do we prime? On distinguishing between semantic priming, procedural priming, and goal priming. In: *The Oxford handbook of human action*, ed. E. Morsella, J. A. Bargh & P. M. Gollwitzer, pp. 173–93. Oxford University Press. [JSU]

Fowers, B. J., Lyons, E., Montel, K. H. & Shaked, N. (2001) Positive illusions about marriage among married and single individuals. *Journal of Family Psychology* 15:95–109. [aJMD]

Frances, B. (2005) When a skeptical hypothesis is live. *Noûs* 39:559–95. [aJMD]

Frank, R. H. (1988) *Passions within reason: The strategic role of the emotions*. Norton. [HV]

Frankfurt, H. (1969) Alternate possibilities and moral responsibility. *Journal of Philosophy* 66:829–39. [HLM]

Frankfurt, H. (1971) Freedom of the will and the concept of a person. *Journal of Philosophy* 68(1):5–20. [MRV]

Frankfurt, H. (1988) *The importance of what we care about*. Cambridge University Press. [SM]

Frankish, K. & Evans, J. St. B. T. (2009) The duality of mind: An historical perspective. In: *In two minds: Dual processes and beyond*, ed. J. Evans & K. Frankish, pp. 1–29. Oxford University Press. [aJMD]

Franks, B. (2011) *Culture and cognition: Evolutionary perspectives*. Palgrave Macmillan. [BF]

Franks, B. (2014) Social construction, evolution and cultural universals. *Culture and Psychology* 20(3):416–39. [BF]

Franks, B., Bangerter, A. & Bauer, M. W. (2013) Conspiracy theories as quasi-religious mentality, an integrated account from cognitive science, social representations theory and frame theory. *Frontiers in Psychology* 4:424. doi: http://dx.doi.org/10.3389/fpsyg.2013.00424. [BF]

Franks, B., Bangerter, A., Bauer, M. W., Hall, M. & Noort, M. C. (2017) Beyond "Monologicality"? Exploring conspiracist worldviews. *Frontiers in Psychology* 8:861. doi: http://dx.doi.org/10.3389/fpsyg.2017.00861. [BF]

Frith, C. D. (2012) The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B, New Thinking: The Evolution of Human Cognition* 367(1599):2213–23. [SB]

Fritsche, I., Kessler, T., Mummendey, A. & Neumann, J. (2009) Minimal and maximal goal orientation and reactions to norm violation. *European Journal of Social Psychology* 39:3–21. doi:10.1002/ejsp.481. [SH]

Frost, P., Nussbaum, G., Loconto, T., Syke, R., Warren, C. & Muise, C. (2013) An individual differences approach to the suggestibility of memory over time. *Memory* 21(3):408–16. [rJMD]

Fumagalli, M. & Priori, A. (2012) Functional and clinical neuroanatomy of morality. *Brain* 135(7):2006–21. [DC]

Funder, D. C. & Ozer, D. J. (1983) Behavior as a function of the situation. *Journal of Personality and Social Psychology* 44(1):107–12. [aJMD]

Funk, C. M. & Gazzaniga, M. S. (2009) The functional brain architecture of human morality. *Current Opinion in Neurobiology* 19(6):678–81. [DC]

Galinsky, A. D., Magee, J. C., Inesi, M. E. & Gruenfeld, D. H. (2006) Power and perspectives not taken. *Psychological Science* 17(12):1068–74. [EL]

Gallagher, S. (2000) Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences* 4(1):14–21. [aJMD]

Gallo, I. S., Keil, A., McCulloch, K. C., Rockstroh, B. & Gollwitzer, P. M. (2009) Strategic automation of emotion regulation. *Journal of Personality and Social Psychology* 96:11–31. [MT]

Gallotti, M., Fairhurst, M. T. & Frith, C. D. (2017) Alignment in social interactions. *Consciousness and Cognition* 48:253–61. [SB]

Gantz, T. & Henkle, G. (2002) *Seatbelts: Current issues.* Prevention Institute. Retrieved from: http://www.preventioninstitute.org/traffic_seatbelt.html. [aJMD]

Gardner, M. P. (1985) Mood states and consumer behavior: A critical review. *Journal of Consumer Research* 12(3):281–300. [ZBW]

Gared, H. J. (1983) Master's liability for the torts of his servant. *The Florida Bar Journal* LVII:597–600. [TAM]

Gavrilets, S. (2012) On the evolutionary origins of the egalitarian syndrome. *Proceedings of the National Academy of Sciences* 109(35):14069–74. [HV]

Gazes, R. P., Hampton, R. R. & Lourenco, S. F. (2015) Transitive inference of social dominance by human infants. *Developmental Science* 20:e12367. doi:10.1111/desc.12367. [BB]

Gelfand, M. J. (2012) Culture's constraints: International differences in the strength of social norms. *Current Directions in Psychological Science* 21:420–24. [BF]

Gibbard, A. (1990) *Wise choices, apt feelings: A theory of normative judgment*. Harvard University Press. [aJMD]

Gibson, J. J. (1977) The theory of affordances. In: *Perceiving, acting and knowing*, ed. R. Shaw & J. Bransford, pp. 67–82. Erlbaum. [BF]

Gigerenzer, G. (2008) Why heuristics work. *Perspectives on Psychological Science* 3:20–29. [NL]

Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. (2016) Comment on estimating the reproducibility of psychological science. *Science* 351:1037-a–38-a. [aJMD]

Gilbert, D. T., Pelham, B. W. & Krull, D. S. (1988) On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology* 54(5):733–40. [aJMD]

Gilbert, M. (1989) *On social facts*. Routledge. [BF]

Gilovich, T. (1991) *How we know what isn't so: The fallibility of human reason in everyday life*. The Free Press. [aJMD]

Gobet, F. & Campitelli, G. (2007) The role of domain-specific practice, handedness, and starting age in chess. *Developmental Psychology* 43(1):159. [rJMD]

Goffman, E. (1958) *The presentation of self in everyday life*. Random House. [DD]

Goffman, E. (1967) *Interaction ritual: Essays on face-to-face behavior*. Anchor. [DD]

Gollwitzer, P. M. (1999) Implementation intentions: Strong effects of simple plans. *American Psychologist* 54:493–503. [MT]

Gollwitzer, P. M. & Sheeran, P. (2006) Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology* 38:69–119. [MT]

Goodwin, G. P. (2015) Moral character in person perception. *Current Directions in Psychological Science* 24:38–44. [LN]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Graham, J., Haidt, J. & Nosek, B. A. (2009) Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96:1029–46. [rJMD, LN]

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S. & Ditto, P. H. (2011) Mapping the moral domain. *Journal of Personality and Social Psychology* 101:366–85. [rJMD, LN]

Granot, Y., Balcetis, E., Schneider, K. E. & Tyler, T. R. (2014) Justice is not blind: Visual attention exaggerates effects of group identification on legal punishment. *Journal of Experimental Psychology: General* 143:2196–208. doi:10.1037/a0037893. [JSU]

Granot, Y., Uleman, J. S. & Balcetis, E. (under review) *The "I" in victim: Just world beliefs and self-relevance as necessary conditions for victim blame*. New York University. [JSU]

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley J. M. & Cohen, J. D. (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293:2105–108. [LN]

Greenwald, A. G., Banaji, M. R. & Nosek, B. A. (2015) Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology* 108(4):553–61. [aJMD]

Haan, N. (1978) Two moralities in action contexts: Relationships to thought, ego regulation, and development. *Journal of Personality and Social Psychology* 36 (3):286–305. [JSU]

Hahlweg, K. & Richtera, D. (2010) Prevention of marital instability and distress. Results of an 11-year longitudinal follow-up study. *Behavior Research and Therapy* 48:377–83. [aJMD]

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

Haidt, J. (2007) The new synthesis in moral psychology. *Science* 316(5827):998–1002. [LN]

Haidt, J. & Joseph, C. (2004) Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* 133:55–66. [SIH]

Hajloo, N., Sadeghi, H., Nadinleoi, K. B. & Habibi, Z. (2014) The role of meta-cognition in students' addiction potential tendency. *International Journal of High Risk Behaviors and Addiction* 3(1):e9355. [JJC]

Hall, L., Johansson, P. & Strandberg, T. (2012) Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS ONE* 7(9): e45457. doi: 10.1371/journal.pone.0045457. [aJMD, EL, SM]

Hall, L., Johansson, P., Tärning, B., Sikström, S. & Deutgen, T. (2010) Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition* 117(1):54–61. [aJMD]

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

Ham, J. & Van den Bos, K. (2008) Not fair for me! The influence of personal relevance on social justice inferences. *Journal of Experimental Social Psychology* 44:699–705. [JSU]

Hamlin, K. J., Wynn, K. & Bloom, P. (2010) Three-month-olds show a negativity bias in their social evaluations. *Developmental Science* 13(6):923–29. [BB]

Hammond, S. I. (2014) Children's early helping in action: Early helping and moral development. *Frontiers in Psychology* 5:1–7. [SIH]

Han, S. & Humphreys, G. (2016) Self-construal: A cultural framework for brain function. *Current Opinion in Psychology* 8:10–14. Available at: http://dx.doi.org/10.1016/j.copsyc.2015.09.013. [BF]

Harman, G. (2000) *Explaining value and other essays in moral philosophy*. Oxford University Press. [aJMD]

Harman, G. H. (1965) The inference to the best explanation. *The Philosophical Review* 74(1):88–95. [rJMD]

Hassin, R. R., Ochsner, K. N. & Trope, Y., eds. (2010) *Self-control in society, mind, and brain (Social cognition and social neuroscience)*. Oxford University Press. [JSU]

Hechler, S. (2016) Cooperation in social groups: Reactions to (moral) deviants (Unpublished doctoral dissertation). Friedrich-Schiller-University of Jena, Jena, Germany. [SH]

Hechler, S., Neyer, F. & Kessler, T. (2016) The infamous among us: Enhanced reputational memory for uncooperative ingroup members. *Cognition* 157:1–13. doi: 10.1016/j.cognition.2016.08.001. [SH]

Hemphill, J. F. (2003) Interpreting the magnitudes of correlation coefficients. *American Psychologist* 58(1):78–80. [aJMD]

Henrich, J., Ensimger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. & J. Ziker (2010) Markets, religion, community size, and the evolution and punishment. *Science* 327:1480–84. [DD]

Henrich, J., Heine, S. J. & Norenzayan, A. (2010) *The weirdest people in the world? Behavioral and Brain Sciences* 33:61–135. [rJMD, LN]

Hill, G. (1982) Group versus individual performance: Are N+1 heads better than one? *Psychological Bulletin* 91:517–39. [DD]

Hirstein, W. (2005) *Brain fiction: Self-deception and the riddle of confabulation*. MIT Press. [aJMD, WH]

Hirstein, W., Sifferd, K. & Fagan, T. (2018) *Responsible brains: Neuroscience and human culpability*. MIT Press. [WH]

Hogarth, R. M. & Einhorn, H. J. (1992) Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology* 24(1):1–55. [ZBW]

Hong, Y-y., Morris, M. W., Chiu, C-y. & Benet-Martinez, V. (2000) Multicultural minds: A dynamic constructivist approach to culture and cognition. *American Psychologist* 55(7):709–20. doi: l0.l037//0003-066X.55.7.709. [JSU]

Horvath, A. O. & Symonds, B. D. (1991) Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology* 38(2):139-159. [aJMD]

Howe, M. J., Davidson, J. W. & Sloboda, J. A. (1998) Innate talents: Reality or myth?. *Behavioral and brain sciences* 21(3):399–407. [rJMD]

Hui, S.-K. A., Wright, R. A., Stewart, C. C., Simmons, A., Eaton, B. & Nolte, R. N. (2009) Performance, cardiovascular, and health behavior effects of an inhibitory strength training intervention. *Motivation and Emotion* 33:419–34. [MT]

Inbar, Y. (2016) Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences* 113(34):E4933–E4934. [aJMD]

Isen, A. M. & Levin, P. F. (1972) Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology* 21(3):384. [AP-C]

Jacoby, L. L. (1991) A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* 30:513–41. [JSU]

Jacquet, J. (2015) *Is shame necessary? New uses for an old tool*. Vintage. [HV]

Jacquet, J., Hauert, C., Traulsen, A. & Milinski, M. (2011) Shame and honour drive cooperation. *Biology Letters* 7:899–901. Available at: https://doi.org/10.1098/rsbl.2011.0367. [HV]

James, W. (1890/1950) *The principles of psychology, vol. 1.* Dover. (Original work published in 1890.) [DD]

Janis, I. L. (1982) *Groupthink*, 2nd edition. Houghton Mifflin. [EL]

Jennings, D., Amabile, T. M. & Ross, L. (1982) Informal covariation assessment: Data-based vs. theory-based judgments. In: *Judgement under uncertainty: Heuristics and biases*, ed. A. Tversky, D. Kahneman & P. Slovic, pp. 211–30. Cambridge University Press. [aJMD]

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

Johansson, P., Hall, L., Sikström, S., Tärning, B. & Lind, A. (2006) How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition* 15(4):673–92. [aJMD]

Johnson, E. J. & Goldstein, D. (2003) Do defaults save lives? *Science* 302 (5649):1338–39. [NL]

Johnson, S. C. (2003) Detecting agents. *Philosophical Transactions of the Royal Society B: Biological Sciences* 358(1431):549–59. [aJMD]

Jones, B., Harris, K. & Tate, W. (2015) Ferguson and beyond: A descriptive epidemiological study using geospatial analysis. *Journal of Negro Education* 84:231–53. [SIH]

Joyce, R. (2006) *The evolution of morality*. MIT Press. [JM]

Kahneman, D. (2011a). Thinking, fast and slow. Farrar, Straus and Giroux. [DD]

Kahneman, D. (2011b). *Thinking, fast and slow*. Penguin. [BF]

Kahneman, D. & Frederick, S. (2002) Representativeness revisited: Attribute substitution in intuitive judgment. In: *Heuristics and biases*, ed. T. Gilovich, D. Griffin & D. Kahneman, pp. 49–81. Cambridge University Press. [aJMD]

Kane, R. (forthcoming). Selfhood, Agency and Responsibility: Reflections on John Doris' *Talking to Our Selves: Reflection, Ignorance, and Agency. Philosophy and Phenomenological Research*. [rJMD]

Kanwisher, N. (2010) Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences* 107(25):11163–70. [DC]

Katsafanas, P. (2013) *Agency and the foundations of ethics: Nietzschean constitutivism*. Oxford University Press. [MA]

Katsafanas, P. (2016) *The Nietzschean self: Moral psychology, agency, and the unconscious*. Oxford University Press. [MA]

Katz, D. (1960) The functional approach to the study of attitudes. Public Opinion Quarterly 24(2):163–204. [DD]

Kawakami, K., Dunn, E., Karmali, F. & Dovidio, J. F. (2009) Mispredicting affective and behavioral responses to racism. Science 323:276–78. [DD]

Keijzer, F. (2013) The Sphex story: How the cognitive sciences kept repeating an old and questionable anecdote. *Philosophical Psychology* 26(4):502–19. [rJMD]

Keltner, D., Van Kleef, G. A., Chen, S. & Kraus, M. W. (2008) A reciprocal influence model of social power: Emerging principles and lines of inquiry. *Advances in Experimental Social Psychology* 40:151–92. [EL]

Keren, G. & Schul, Y. (2009) Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science* 4(6):533–50. [aJMD]

Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M. & Nelson, C. B. (1995) Posttraumatic stress disorder in the National Comorbidity Survey. *Archives of General Psychiatry* 52:1048–60. [LN]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Kessler, T. & Cohrs, J. C. (2008) The evolution of authoritarian processes: How to commit group members to group norms. *Group Dynamics Theory, Research, and Practice* 12:73–84. doi:10.1037/1089-2699.12.1.73. [SH]

Kessler, T., Neumann, J., Mummendey, A., Berthold, A., Schubert, T. & Waldzus, S. (2010) How do we assign punishment? The impact of minimal and maximal standards on the evaluation of deviants. *Personality and Social Psychology Bulletin* 36:1213–24. doi:10.1177/0146167210380603. [SH]

Khader, S. (2011) *Adaptive preferences and women's empowerment*. Oxford University Press. [MRV]

Kiehl, K. (2014) *The psychopath whisperer*. Oneworld. [rJMD]

King's Bench (1466) Y.B.M. 6 Edw. IV, folio 7, placitum 18. Available at: http://www.casebriefs.com/blog/law/torts/torts-keyed-to-prosser/development-of-liability/hulle-v-orynge-the-case-of-thorns/ [accessed on 6 February 2017]. [TAM]

Kitayama, S. & Imada, T. (2008) Defending cultural self. In: *Advances in motivation and achievement, vol. 15: Social Psychological Perspectives*, ed. M. L. Maehr, S. A. Karabenick & T. C. Urdan, pp. 171–208. JAI Press. [BF]

Kiverstein, J., ed. (2016) *The Routledge handbook of philosophy of the social mind*. Routledge. [SB]

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. & Damasio, A. (2007) Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446(7138):908–11. [LN]

Kornblith, H. (2010) What reflective endorsement cannot do. *Philosophy and Phenomenological Research* 80(1):1–19. [aJMD]

Kornblith, H. (2012) *On reflection*. Oxford University Press. [aJMD]

Korsgaard, C. M. (1996) *The sources of normativity*. Cambridge University Press. [aJMD]

Korsgaard, C. M. (2009) *Self-constitution: Agency, identity, and integrity*. Oxford University Press. [aJMD]

Kristjánsson, K. (2013) *Virtues and vices in positive psychology: A philosophical critique*. Cambridge University Press. [BJF]

Krosnick, J. A., Miller, J. M. & Tichy, M. P. (2004) An unrecognized need for ballot reform: Effects of candidate name order. In: *Rethinking the vote: The politics and prospects of American election reform*, ed. A. N. Crigler, M. R. Just & E. J. McCaffery, pp. 51–74. Oxford University Press. [aJMD, ZBW]

Kruger, J., Wirtz, D. & Miller, D. T. (2005) Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology* 88(5):725–35. [JJC]

Kruglanski, A. W. (1996) Motivated social cognition: Principles of the interface. In: *Social psychology: Handbook of basic principles*, ed. E. T. Higgins & A. W. Kruglanski, pp. 493–520. Guilford Press. [aJMD]

Krupnick, J. L., Sotsky, S. M., Simmens, S., Moyher, J., Elkin, I., Watkins, J. & Pilkonis, P. A. (1996) The role of the therapeutic alliance in psychotherapy and pharmacotherapy outcome: Findings in the National Institute of Mental Health Treatment of Depression Collaborative Research Project. *Journal of Consulting and Clinical Psychology* 64:532–39. [aJMD]

Kumar, V. & May, J. (under review) How to debunk moral beliefs. [JM]

Kunda, Z. (1990) The case for motivated reasoning. *Psychological Bulletin* 108:480–98. [aJMD]

Lambert, M. J. & Ogles, B. M. (2004) The efficacy and effectiveness of psychotherapy. In: *Bergin and Garfield's handbook of psychotherapy and behavior change*, ed. M. J. Lambert, pp. 139–93. Wiley. [aJMD]

Landy, J. F. & Goodwin, G. P. (2015) Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science* 10(4):518–36. [JM]

Langer, E. J. (1975) The illusion of control. *Journal of Personality and Social Psychology* 32:311–28. [aJMD]

Latané, B. & Nida, S. (1981) Ten years of research on group size and helping. *Psychological Bulletin* 89(2):308–24. [JM]

Lefevor, G. T. & Fowers, B. J. (2016) Traits, situational factors, and their interactions as explanations of helping behavior. *Journal of Personality and Individual Differences* 92:159–63. doi.org/10.1016/j.paid.2015.12.042. [BJF]

Lefevor, G. T., Fowers, B. J., Ahn, S., Lang, S. F. & Cohen, L. M. (2017) To what degree do situational influences explain spontaneous helping behaviour? A meta-analysis. *European Review of Social Psychology* 28:227–56. Available at: http://dx.doi.org/10.1080/10463283.2017.1367529. [BJF]

Levy, N. (2011) *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press. [aJMD, HV]

Levy, N. (2014) *Consciousness and moral responsibility*. Oxford University Press. [NL]

Liberman, V., Samuels, S. M. & Ross, L. (2004) The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. Personality and Social Psychology, Bulletin 30:1175–85. [DD]

Lindemann, H. (2014) *Holding and letting go: The social practice of personal identities*. Oxford University Press. [MA]

Lindenberg, S. (2012) How cues in the environment affect normative behavior. In: *Environmental psychology: An introduction*, ed. L. Steg, A. E. van denBerg & J. I. M. de Groot, pp. 119–28. Wiley. [DD]

List, J. A. (2007) On the interpretation of giving in dictator games. *Journal of Political Economy* 115:482–93. [DD]

Lonergan, K. (director) (2016) *Manchester by the sea*. Amazon Studios. [TS]

Luborsky, L., McLellan, A. T., Woody, G. E., O'Brien, C. P. & Auerbach, A. (1985) Therapist success and its determinants. *Archives of General Psychiatry* 42:602–11. [aJMD]

Luborsky, L. & Singer, B. (1975) Comparative studies of psychotherapies: Is it true that "Everyone has won and all must have prizes"? *Archives of General Psychiatry* 32:995–1008. [aJMD]

Luetchford, M. (2001) Right and wrong in Buddhism [transcript]. Available at: http://www.dogensangha.org.uk/PDF/rightwrong.pdf. [DC]

Luhrmann, T. M. (2011) Hallucinations and sensory overrides. *Annual Review of Anthropology* 40:71–85. [DC]

Lutz, G. (2010) First come, first served: The effect of ballot position on electoral success in open ballot PR elections. *Representation* 46:167–81. [aJMD]

Machery, E. (2009) *Doing without concepts*. Oxford University Press. [aJMD]

Machery, E. & Doris, J. M. (2017) An open letter to our students: Doing interdisciplinary moral psychology. In: *Moral psychology: A multidisciplinary guide*, ed. B. G. Voyer & T. Tarantola, pp. 119–43. Springer. [arJMD]

Machery, E. & Mallon, R. (2010) Evolution of morality. In: *The moral psychology handbook*, ed. J. M. Doris and the Moral Psychology Research Group, pp. 3–46. Oxford University Press. [rJMD]

Mackie, J. L. (1977) *Ethics: Inventing right and wrong*. Penguin. [HV]

Maibom, H. L. (2014) Knowing what we are doing. In: *Moral psychology and human agency*, ed. J. D'Arms & D. Jacobson, pp. 108–22. Oxford University Press. [HLM]

Malle, B. F. (2006) The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin* 132(6):25. [HM]

Marcinkiewicz, K. (2014) Electoral contexts that assist voter coordination: Ballot position effects in Poland. *Electoral Studies* 33:322–34. [aJMD, NL]

Markus, H. R. & Kitayama, S. (1991) Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review* 98:224–53. [BF, JSU]

Markus, H. R. & Kitayama, S. (2003) Models of agency: Sociocultural diversity in the construction of action. In: *The 49th Annual Nebraska Symposium on Motivation: Cross-Cultural Differences in Perspectives on the Self*, ed. G. Berman & J. Berman, pp. 2–57. University of Nebraska Press. [BF]

Markus, H. R. & Kitayama, S. (2010) Cultures and selves: A cycle of mutual constitution. *Perspectives on Psychological Science* 5(4):420–30. [BF]

Marotta, A., Tinazzi, M., Cavedini, C., Zampini, M. & Fiorio, M. (2016) Individual differences in the rubber hand illusion are related to sensory suggestibility. *PLoS One* 11(12):e0168489. [rJMD]

Marques, J. M., Abrams, D., Paez, D. & Martinez-Taboada, C. (1998) The role of categorization and in-group norms in judgments of groups and their members. *Journal of Personality and Social Psychology* 75(4):976–88. doi:10.1037/0022-3514.75.4.976. [SH]

Martin, D. J., Garske, J. P. & Davis, M. K. (2000) Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology* 68(3):438. [aJMD]

Maruna, S. & Copes, H. (2005) What have we learned in five decades of neutralization research? *Crime and Justice* 32:221–320. doi:10.1086/655355. [PMP]

Mascaro, O. & Csibra, G. (2012) Representation of stable social dominance relations by human infants. *Proceedings of the National Academy of Sciences* 109 (18):6862–67. [BB]

May, J. (2013) Skeptical hypotheses and moral skepticism. *Canadian Journal of Philosophy* 43(3):341–59. [JM]

May, J. (2014) Does disgust influence moral judgment? *Australasian Journal of Philosophy* 92(1):125–41. [JM]

May, J. (forthcoming). *Regard for reason in the moral mind*. Oxford University Press. [JM]

McKenna, M. (2017) Reasons-responsive theories of freedom. In: *The Routledge companion to free will*, ed. K. Timpe, M. Griffith & N. Levy, pp. 27–40. Routledge. [NL]

McKenna, M. & Pereboom, D. (2016) *Free will: A contemporary introduction*. Routledge. [MRV]

Meindl, P., Jayawickreme, E., Furr, R. M. & Fleeson, W. (2013) A foundation beam for studying morality from a personological point of view: Are individual differences in moral behaviors and thoughts consistent? *Journal of Research in Personality* 59:81–92. doi.org/10.1016/j.jrp.2015.09.005. [BJF]

Mele, A. (2009) Moral responsibility and history revisited. *Ethical Theory and Moral Practice* 12(5):463–75. [MRV]

Mendez, M. F. (2009) The neurobiology of moral behavior: Review and neuropsychiatric implications. *CNS Spectrums* 14(11):608–20. [DC]

Mendoza, S. A., Gollwitzer, P. M. & Amodio, D. M. (2010) Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin* 36:512–23. [MT]

Mercier, H. & Sperber, D. (2009) Intuitive and reflective inferences. In: *In two minds: Dual processes and beyond*, ed. J. St. B. T. Evans & K. Frankish, pp. 149–70. Oxford University Press. [rJMD]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Mercier, H. & Sperber, D. (2011) Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34:57–111. doi:10.1017/S0140525X10000968. [PMP, JSU]

Mercier, H. & Sperber, D. (2017) *The enigma of reason*. Harvard University Press. [rJMD, HM]

Meredith, M. & Salant, Y. (2013) On the causes and consequences of ballot order effects. *Political Behavior* 35(1):175–97. [aJMD]

Metcalfe, J. & Mischel, W. (1999) A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review* 106(1):3. [aJMD]

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman E. J., Kubiszyn, T. W. & Reed, G. M. (2001) Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist* 56(2):128–65. [aJMD]

Miles, J. B. (2015) *The free will delusion: How we settled for the illusion of morality*. Troubador. [PMP]

Milgram, S. (1974) *Obedience to authority: An experimental view*. Harper & Row. [SIH, EL]

Miller, C. B. (2013) *Moral character*. Oxford University Press. [JM]

Miller, J. G. (1984) Culture and the development of everyday social explanation. *Journal of Personality and Social Psychology* 46(5):961–78. [JSU]

Miller, R. L., Brickman, P. & Bolen, D. (1975) Attribution versus persuasion as a means for modifying behavior. *Journal of Personality and Social Psychology* 31 (3):430–41. doi:10.1037/h0076539. [SH]

Miller, R. M., Hannikainen, I. A. & Cushman, F. A. (2014) Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion* 14(3):573–87. [LN]

Millikan, R. G. (2000) *On clear and confused ideas: An essay on substance concepts*. *Cambridge studies in philosophy*. Cambridge University Press. doi:10.1017/CBO9780511613296. [SH]

Mischel, M., Ebbesen, E. B. & Zeiss, A. R. (1972) Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology* 21:204–18. [BJF]

Mocellin, R., Walterfang, M. & Velakoulis, D. (2006) Neuropsychiatry of complex visual hallucinations. *Australian and New Zealand Journal of Psychiatry* 40 (9):742–51. [DC]

Monroe, A. E. & Malle, B. F. (2010) From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology* 1:211–24. doi:10.1007/s13164-009-0010-7. [PMP]

Montero, B. (2010) Does bodily awareness interfere with highly skilled movement? *Inquiry* 53(2):105–22. [aJMD]

Mooijman, M., Meindl, P., Oyserman, D., Dehghani, M., Monteresso, J., Doris, J. M. & Graham, J. (forthcoming) Resisting temptation for the good of the group: Binding moral values and the moralization of self-control. *Journal of Personality and Social Psychology*. [rJMD]

Moran, R. (2001) *Authority and estrangement: An essay on self-knowledge.* Princeton University Press. [aJMD]

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H. & Kassam, K. S. (2015) Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences* 2 (1):129–40. [MT]

Morris, I. (2015) *Foragers, farmers, and fossil fuels: How human values evolve*. Princeton University Press. [HV]

Morse, S. J. (2008) Psychopathy and criminal responsibility. *Neuroethics* 1(3):205–12. [rJMD]

Moscovici, S. & Zavalloni, M. (1969) The group as a polarizer of attitudes. *Journal of Personality and Social Psychology* 12:125–35. [EL]

Muraven, M., Baumeister, R. F. & Tice, D. M. (1999) Longitudinal improvement of self regulation through practice: Building self-control strength through repeated exercise. *Journal of Social Psychology* 139:446–57. [MT]

Murray, S. (2017) Responsibility and vigilance. *Philosophical Studies* 174(2):507–27. [SM]

Nahmias, E. (2011) Intuitions about free will, determinism, and bypassing. In: *Oxford handbook of free will*, 2nd edition, ed. R. Kane, pp. 555–76. Oxford University Press. [arJMD]

Nelkin, D. (2011) *Making sense of freedom and responsibility*. Oxford University Press. [aJMD]

Nelkin, D. K. (forthcoming). Responsibility and ignorance of the self: Comments on John Doris' *Talking to Our Selves: Reflection, Ignorance, and Agency*. *Social Theory and Practice*. [rJMD]

Nelkin, D. K. (2008) Responsibility and rational abilities: Defending an asymmetrical view. *Pacific Philosophical Quarterly* 89(4):497–515. [SM]

Nelson, T. O. & Narens, L. (1990) Metamemory: A theoretical framework and new findings. In: *The Psychology of learning and motivation: Advances in research and theory*, ed. G. Bower, pp. 125–73. Academic Press. [JJC]

Nichols, S. (2007) After incompatibilism: A naturalistic defense of the reactive attitudes. *Philosophical Perspectives* 21(1):405–28. [aJMD]

Nichols, S. (2014) Process debunking and ethics. *Ethics* 124:727–49. [JM]

Niemi, L. & Young, L. (2016) When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin* 42(9):1227–42. [rJMD, LN]

Nietzsche, F. (1878/1996) *Human, all too human,* trans. R. J. Hollingdale. Cambridge University Press. (Original work published in 1878.) [MA]

Nietzsche, F. (1881/1997) *Daybreak: Thoughts on the prejudices of morality,* trans. R. J. Hollingdale, ed. M. Clark & B. Leiter. Cambridge University Press. (Original work published in 1881.) [MA]

Nietzsche, F. (1882/2001) *The gay science,* trans. J. Nauckhoff, ed. B. Williams. Cambridge University Press. (Original work published in 1882.) [MA]

Nietzsche, F. (1883/2006) *Thus spoke Zarathustra*, trans. A. Del Caro, ed. A. Del Caro & R. Pippin. Cambridge University Press. (Original work published in 1883.) [MA]

Nietzsche, F. (1886/2001) *Beyond good and evil*, trans. J. Norman, ed. R.-P. Horstmann. Cambridge University Press. (Original work published in 1886.) [MA]

Nietzsche, F. (1887/2006) *Genealogy of morals,* trans. C. Diethe, ed. K. Ansell-Pearson. Cambridge University Press. (Original work published in 1887.) [MA]

Nisbett, R. & Wilson, T. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84:231–59. [DD]

Nisbett, R. E., Peng, K., Choi, I. & Norenzayan, A. (2001) Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review* 108(2):291–310. doi: 10.1037//0033-295X.108.2.291. [JSU, ZBW]

Nizzi, M-C., Demertzi, A., Gosseries, O., Bruno, M-A., Jouen, F. & Laureys, S. (2012) From armchair to wheelchair: How patients with a locked-in syndrome integrate bodily changes in experienced identity. *Consciousness and Cognition* 21:431–37. [LN]

Nizzi, M-C. & Niemi, L. (in preparation) The sense of self predicts suicidal thoughts and behaviors in survivors of sexual assault. [rJMD, LN]

Noë, A. (2012) *Varieties of presence*. Harvard University Press. [aJMD]

Noonan, H. (1989) *Personal identity*. Routledge. [aJMD]

Norman, J. (2002) Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches. *Behavioral and Brain Sciences* 25(1):73–96. [aJMD]

Northover, S. B., Pedersen, W. C., Cohen, A. B. & Andrews, P. W. (2017) Artificial surveillance cues do not increase generosity: Two meta-analyses. *Evolution and Human Behavior* 38(1):144–53. [ZBW]

Null, J. (2016) *Heatstroke deaths of children in vehicles*. Available at: http://noheatstroke.org. [SM]

Olin, L. & Doris, J. M. (2014) Vicious minds. *Philosophical Studies* 168(3):665–92. [MT]

O'Malley, M. & Andrews, L. (1983) The effect of mood and incentives on helping: Are there some things money can't buy? *Motivation and Emotion* 7:179–89. [MT]

Onu, D., Smith, J. & Kessler, T. (2015) Intergroup emulation: An improvement strategy for lower-status groups. *Group Processes & Intergroup Relations* 18:210–24. doi: 10.1177/1368430214556698. [SH]

Onu, D., Kessler, T. & Smith, J. (2016a) Admiration: A conceptual review of the knowns and unknowns. *Emotion Review* 8(3):1–13. doi:10.1177/1754073915610438. [SH]

Onu, D., Kessler, T., Smith, J. R., Andnovskia-Trajkovska, Fritsche, I., Midson, G. R. & Smith, J. R. (2016b) Inspired by the outgroup: A social identity analysis of intergroup admiration. *Group Processes & Intergroup Relations* 19:713–31. doi:10.1177/1368430216629811. [SH]

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716. [aJMD]

Orlinsky, D. E., Ronnestad, M. H. & Willutzki, U. (2004) Fifty years of process-outcome research: Continuity and change. In: *Bergin and Garfield's handbook of psychotherapy and behavior change*, 5th edition, ed. M. J. Lambert, pp. 307–89. Wiley. [aJMD]

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J. & Tetlock, P. E. (2013) Predicting ethnic and racial discrimination. *Journal of Personality and Social Psychology* 105(2):171–92. [JM]

Overton, W. F. (2006) Developmental psychology: Philosophy, concepts, methodology. In: *Theoretical models of human development, vol. 1 of the Handbook of child psychology*, 6th edition, ed. R. M. Lerner, pp. 18–88. Wiley. [SIH]

Park, G., Kappes, A., Rho, Y. & Van Bavel, J. J. (2016) At the heart of morality lies neuro-visceral integration: Lower cardiac vagal tone predicts utilitarian moral judgment. *Social Cognitive and Affective Neuroscience* 11(10):1588–96. [LN]

Parks-Stamm, E. J. & Gollwitzer, P. (2009) Goal implementation: The benefits and costs of IF-THEN planning. In: *The psychology of goals*, ed. G. B. Moskowitz & H. Grant, pp. 362–91. Guilford Press. [MT]

Pasek, J., Schneider, D., Krosnick, J. A., Tahk, A., Ophir, E. & Milligan, C. (2014) Prevalence and moderators of the candidate name-order effect evidence from statewide general elections in California. *Public Opinion Quarterly* 78(2):416–39. [BJF]

Pauer-Studer, H. & Velleman, D. J. (2011) Distortions of normativity. *Ethical Theory and Moral Practice* 14:329–56. doi:10.1007/s10677-010-9246-7. [SH]

Peck, R. (director) (2016) *I am not your Negro* [motion picture]. Velvet Film. [BB]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Pereboom, D. (2014) *Free will, agency, and meaning in life*. Oxford University Press. [rJMD]

Perkins, A. M., Leonard, A. M., Weaver, K., Dalton, J. A., Mehta, M. A., Kumari, V., Williams, S. C. R. & Ettinger, U. (2013) A dose of ruthlessness: Interpersonal moral judgment is hardened by the anti-anxiety drug lorazepam. *Journal of Experimental Psychology: General* 142(3):612–20. [LN]

Peters, K. & Kashima, Y. (2007) From social talk to social action: Shaping the social triad with emotion sharing. *Journal of Personality and Social Psychology* 93(5):780–97. doi:10.1037/0022-3514.93.5.780. [SH]

Piaget, J. (1932/1965) *The moral judgment of the child*, trans. M. Gabain. Free Press. (Original work published 1932.) [SIH]

Piaget, J. (1936/1963) *The origins of intelligence in children*, trans. M. Cook. Norton. (Original work published in 1936.) [SIH]

Piaget, J. (1974/1976) *The grasp of consciousness: Action and concept in the young child*. Harvard University Press. (Original work published in 1974.) [rJMD, SIH]

Pinker, S. (2011) *The better angels of our nature: Why violence has declined*. Viking Books. [HV]

Pinkerton v. United States (1946) 328 U.S. 640. [TAM]

Pinto, I. R., Marques, J. M., Levine, J. M. & Abrams, D. (2010) Membership status and subjective group dynamics: Who triggers the black sheep effect? *Journal of Personality and Social Psychology* 99(1):107–19. doi:10.1037/a0018187. [SH]

Plant, E. A. & Peruche, B. M. (2005) The consequences of race for police officers' responses to criminal suspects. *Psychological Science* 16:180–83. [MT]

Plant, E. A., Peruche, B. M. & Butz, D. A. (2005) Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental and Social Psychology* 41:141–56. [MT]

Prinz, J. J. (2004) *Gut reactions: A perceptual theory of emotion*. Oxford University Press. [aJMD]

Railton, P. (2003) *Facts, values, and norms: Essays toward a morality of consequence*. Cambridge University Press. [aJMD]

Railton, P. (2006) How to engage reason: The problem of regress. In: *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. R. J. Wallace, P. Pettit, S. Scheffler & M. Smith, pp. 176–201. Oxford University Press. [ZBW]

Railton, P. (2009) Practical competence and fluent agency. In: *Reasons for action*, ed. D. Sobel & S. Wall, pp. 81–115. Cambridge University Press. [ZBW]

Ramachandran, V. S. (1996) What neurological syndromes can tell us about human nature: Some lessons from phantom limbs, Capgras' syndrome, and anosognosia. In: *Cold Spring Harbor symposia on quantitative biology, vol. 61*, pp. 115–34. Cold Spring Harbor Laboratory Press. [aJMD]

Raz, J. (1999) *Engaging reason: On the theory of value and action*. Oxford University Press. [ZBW]

Reagan, R. (1987) Iran arms and Contra aid controversy. PBS. Available at: http://www.pbs.org/wgbh/americanexperience/features/primary-resources/reagan-iran-contra/. [TAM]

Reber, A. S. (1993) *Implicit learning and tacit knowledge*. Oxford University Press. [aJMD]

Reddy, V., Liebal, K., Hicks, K., Jonnalagadda, S. & Chintalapuri, B. (2013) The emergent practice of infant compliance. An exploration in two cultures. *Developmental Psychology* 49(9):1754–62. [JZ]

Rees, G. (2014) Hallucinatory aspects of normal vision. In: *The neuroscience of visual hallucinations*, ed. D. Collerton, E. Perry & U. P. Mosimann, pp. 47–57. Wiley. [DC]

Robinson, P. H. & Grall, J. A. (1983) Element analysis in defining criminal liability: The model penal code and beyond. *Stanford Law Review* 35:681–762. [TAM]

Rochat, P., Dias, M. D., Liping, G., Broesch, T., Passos-Ferreira, C., Winning, A. & Berg, B. (2009) Fairness in distributive justice by 3- and 5-year-olds across seven cultures. *Journal of Cross-Cultural Psychology* 40(3):416–42. [BB]

Roediger, H. L. & Butler, A. C. (2011) Paradoxes of learning and memory. In: *The paradoxical brain*, ed. N. Kapur, pp. 151–76. Cambridge University Press. [rJMD]

Roediger, H. L. III (1990) Implicit memory: Retention without remembering. *American Psychologist* 45:1043–56. [aJMD]

Roediger, H. L., III & Karpicke, J. D. (2006) Test-enhanced learning. *Psychological Science* 17(3):249–55. [JJC]

Rosenstein, D. & Oster, H. (1988) Differential facial responses to four basic tastes in newborns. *Child Development* 59:1555–68. [BB]

Rossi, G. (2012) Bilateral and unilateral requests: The use of imperatives and mi X? interrogatives in Italian. *Discourse Processes* 49(5):426–58. [JZ]

Rutland, A. & Killen, M. (2015) A developmental science approach to reducing prejudice and social exclusion. *Social Issues and Policy Review* 9:121–54. [SIH]

Sarason, I., Smith, R. & Diener, E. (1975) Personality research: Components of variance attributable to the person and the situation. *Journal of Personality and Social Psychology* 32:199–204. [SIH]

Sasaki, J. Y. & Kim, H. S. (2017) Nature, nurture, and their interplay. *Journal of Cross-Cultural Psychology* 48(1):4–22. doi:10.1177/0022022116680481. [BF]

Sassenberg, K., Moskowitz, G. B., Fetterman, A. & Kessler, T. (2017) Priming creativity as a strategy to increase creative performance by facilitating the activation and use of remote associations. *Journal of Experimental Social Psychology* 68:128–38. doi:10.1016/j.jesp.2016.06.010. [SH]

Savani, K., Markus, H. R. & Conner, A. L. (2008) Let your preference by your guide? Preferences and choices are more tightly linked for North Americans and for Indians. *Journal of Personality and Social Psychology* 95:861–76. [DD]

Savani, K., Markus, H. R., Naidu, N. V. R., Kumar, S. & Berlia, V. (2010) What counts as a choice? U.S. Americans are more likely than Indians to construe actions as choices. *Psychological Science* 21:391–98. [DD]

Sayers, S. L., Kohn, C. S. & Heavey, C. (1998) Prevention of marital dysfunction: Behavioral approaches and beyond. *Clinical Psychology Review* 18(6):713–44. [aJMD]

Scanlon, T. (1998) *What we owe to each other*. Harvard University Press. [SB]

Scanlon, T. (2013) Interpreting blame. In: *Blame. Its nature and norms*, ed. J. Coates & N. Tognazzini, pp. 84–99. Oxford University Press. [HV]

Schechtman, M. (1996) *The constitution of selves*. Cornell University Press. [aJMD]

Schechtman, M. (2011) The narrative self. In: *The Oxford handbook of the self*, ed. S. Gallagher, pp. 394–416. Oxford University Press. [aJMD]

Schneider, W. & Shiffrin, R. M. (1977) Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* 84(1):1–66. [aJMD]

Schwartz, D. (1995) The emergence of abstract representations in dyad problem solving. *Journal of the Learning Sciences* 4(3):321–54. [DD]

Schwarz, N., Sanna, L. J., Skurnik, I. & Yoon, C. (2007) Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology* 39:127–61. [JJC]

Schwartz, S. H. & Sagie, G. (2000) Value consensus and importance: A cross-national study. *Journal of Cross-Cultural Psychology* 31:465–97. [BF]

Scott, M. B. & Lyman, S. M. (1968) Accounts. *American Sociological Review* 33 (1):46–62. [PMP]

Searle, J. (1995) *The construction of social reality*. Penguin. [BF]

Seligman, M. E. P. (1993) *What you can change and what you can't: The complete guide to successful self-improvement*. Knopf. [aJMD]

Seligman, M. E. P., Railton, P., Baumeister, R. F. & Sripada, C. S. (2016) *Homo prospectus*. Oxford University Press. [JM]

Selke, S. (2016) *Lifelogging: Digital self-tracking and lifelogging: Between disruptive technology and cultural transformation*. Springer. [MA]

Sellaro, R., Nitsche, M. A. & Colzato, L. S. (2016) The stimulated social brain: Effects of transcranial direct current stimulation on social cognition. *Annals of the New York Academy of Sciences* 1369(1):218–39. [DC]

Shalvi, S., Gino, F., Barkan, R. & Ayal, S. (2015) Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science* 24 (2):125–30. doi:10.1177/0963721414553264. [PMP]

Sherif, M., Harvey, O. J., White, B. J., Hood, W. & Sherif, C. (1961) *Intergroup conflict and cooperation: The robbers cave experiment*. University of Oklahoma Institute of Group Relations. [EL]

Shiffrin, R. M. & Schneider, W. (1977) Controlled and automatic human information processing. II. Perceptual learning, automatic attending and a general theory. *Psychological Review* 84:127–90. [MT]

Shimizu, Y., Lee, H. & Uleman, J. S. (2017) Culture as automatic processes for making meaning: Spontaneous trait inferences. *Journal of Experimental Social Psychology* 69(1):79–85. doi: 10.1016/j.jesp.2016.08.003. [rJMD, JSU]

Shoemaker, D. (2015). Review of *Talking to Our Selves: Reflection, Ignorance, and Agency*, by John M. Doris. Notre Dame Philosophical Reviews, 2015.11.17. Available at: http://ndpr.nd.edu/news/62138-talking-to-our-selves-reflection-ignorance-and-agency/. [rJMD, AP-C]

Shults, R. A., Elder, R. W., Sleet, D. A., Thompson, R. S. & Nichols, J. L. (2004) Primary enforcement seat belt laws are effective even in the face of rising belt use rates. *Accident Analysis & Prevention* 36(3):491–93. [aJMD]

Siegel, J. Z. & Crockett, M. J. (2013) How serotonin shapes moral judgment and behavior. *Annals of the New York Academy of Sciences* 1299(1):42–51. [DC]

Simon, J. (1998) An analysis of the distribution of combinations chosen by UK national lottery players. *Journal of Risk and Uncertainty* 17(3):243–77. [aJMD]

Simons, D. J. & Chabris, C. F. (1999) Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* 28:1059–74. [JJC]

Simons, D. J. & Rensink, R. A. (2005) Change blindness: Past, present, and future. *Trends in cognitive sciences* 9(1):16–20. [DC]

Sinnott-Armstrong, W. (2006) *Moral skepticisms*. Oxford University Press. [JM]

Sloman, S. A. (1996) The empirical case for two systems of reasoning. Psychological Bulletin 119:3–22. [DD]

Smart, J. J. C. (1961) Free will, praise, and blame. *Mind* 70:291–306. [rJMD]

Smilansky, S. (2000) *Free will and illusion*. Oxford University Press. [PMP]

Smith, A. M. (2005) Responsibility for attitudes: Activity and passivity in mental life. *Ethics* 115(2):236–71. [aJMD]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Smith, J. D., Couchman, J. J. & Beran, M. J. (2012) The highs and lows of theoretical interpretation in animal-metacognition research. *Philosophical Transactions of the Royal Society B: Biological Sciences* 397:1297–309. [JJC]

Snyder, D. K., Castellani, A. M. & Whisman, M. A. (2006) Current status and future directions in couple therapy. *Annual Review of Psychology* 57:317–44. [aJMD]

Sokol, B. W., Hammond, S. I., Kuebli, J. & Sweetman, L. (2015) The development of agency. In: *Handbook of child psychology and developmental science*, 7th edition, ed. W. F. Overton & P. C. M. Molenaar, pp. 284–322. Wiley. [SIH]

Sosa, E. (1991) Knowledge and intellectual virtue. In: *Knowledge in perspective: Essays in epistemology*, ed. E. Sosa, pp. 225–44. Cambridge University Press. [rJMD]

Sparks, A. & Barclay, P. (2013) Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behavior* 34(5):317–22. [JM]

Sripada, C. (2015a) Moral responsibility, reasons, and the self. *Oxford Studies in Agency and Responsibility* 3:242–64. [aJMD]

Sripada, C. (2015b) Self-expression: A deep self theory of moral responsibility. *Philosophical Studies* 175(5):1203–232. [aJMD, MRV]

Sripada, C. & Stich, S. (2006) A framework for the psychology of norms. In: *The innate mind, vol. 2: Culture and cognition*, ed. P. Carruthers, S. Laurence & S. P. Stich, pp. 280–301. Oxford University Press. [rJMD]

Stanovich, K. E. (2004) *The robo's rebellion: Finding meaning in the age of Darwin*. University of Chicago Press. [aJMD]

Stanovich, K. E. (2011) *Rationality and the reflective mind*. Oxford University Press. [BF]

Stepanikova, I., Triplett, J. & Simpson, B. (2011) Implicit racial bias and prosocial behavior. *Social Science Research* 40:1186–95. [MT]

Steptoe, A., Wardle, J. & Marmot, M. (2005) Positive affect and health-related neuroendocrine, cardiovascular, and inflammatory processes. *Proceedings of the National Academy of Sciences of the United States of America* 102(18):6508–12. [aJMD]

Stich, S. P. (1990) *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation*. MIT Press. [DD]

Stivers, T., Mondada, L. & Steensig, J., eds. (2011) *The morality of knowledge in conversation*. Cambridge University Press. [JZ]

Strawson, G. (1994) The impossibility of moral responsibility. *Philosophical Studies* 75(1):5–24. doi:10.1007/BF00989879. [PMP]

Strawson, G. (2004) Against narrativity. *Ratio* 17(4):428–52. [aJMD]

Strawson, P. F. (1962) Freedom and resentment. *Proceedings of the British Academy* 48:1–25. [SB, arJMD, SH, PMP, TS]

Sunstein, C. (2009) *Going to extremes: How like minds unite and divide*. Oxford University Press. [EL]

Swann, W. B., Jr. & Buhrmester, M. (2015) Identity fusion. *Current Directions in Psychological Science* 24:52–57. [BF]

Swim, J. K. & Hyers, L. L. (1999) Excuse me—what did you just say?! Women's public and private responses to sexist remarks. *Journal of Experimental Social Psychology* 35:68–88. [DD]

Sykes, G. M. & Matza, D. (1957) Techniques of neutralization: A theory of delinquency. *American Sociological Review* 22(6):664–70. [PMP]

Tajfel, H., Billig, M. G., Bundy, R. P. & Flament, C. (1971) Social categorization and intergroup behavior. *European Journal of Social Psychology* 1:149–77. [EL]

Tate, W. F., IV (2016) "Dream by her river": Ferguson, Missouri, and the geography of opportunity. *Paper presented at the Jean Piaget Society Meeting, Chicago, IL.* [SIH]

Taylor, C. (1994) The politics of recognition. In: *Multiculturalism: Examining the politics of recognition*, ed. A. Gutmann, pp. 25–73. Princeton University Press. [aJMD]

Taylor, S. E., Lerner, J. S., Sherman, D. K., Sage, R. M. & McDowell, N. K. (2003) Are self-enhancing cognitions associated with healthy or unhealthy biological profiles? *Journal of Personality and Social Psychology* 85(4):605–15. [aJMD]

Tetlock, P. (2009) *Expert political judgment: How good is it? How can we know?* Princeton University Press. [MT]

Thompson, S. C., Sobolew-Shubin, A., Galbraith, M. E., Schwankovsky, L. & Cruzen, D. (1993) Maintaining perceptions of control: Finding perceived control in low-control circumstances. *Journal of Personality and Social Psychology* 64(2):293–304. [aJMD]

Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M. & Carey, S. (2011) Big and mighty: Preverbal infants mentally represent social dominance. *Science* 331 (6016):477–80. [BB]

Thomson, C., Wilson, R., Collerton, D., Freeston, M. & Dudley, R. (2017) Cognitive behavioural therapy for visual hallucinations: An investigation using a single-case experimental design. *The Cognitive Behaviour Therapist* 10. Available at: https://doi.org/10.1017/S1754470X17000174. [DC]

Tiberius, V. (2002) Virtue and practical deliberation. *Philosophical Studies* 111 (2):147–72. [aJMD, DD]

Tiberius, V. (forthcoming) Comments on John Doris, *Talking to ourselves: Reflection, ignorance, and agency*. Philosophy and Phenomenological Research. [rJMD]

Todd, P. M. & Gigerenzer, G. (2007) Environments that make us smart ecological rationality. *Current Directions in Psychological Science* 16:167–71. [NL]

Todorov, A. & Uleman, J. S. (2004) The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology* 87:482–93. [JSU]

Tomasello, M. (2009) *Why we cooperate*. MIT Press. [BF]

Tomasello, M. (2016) *A natural history of human morality*. Harvard University Press. [SB]

Toneatto, T., Blitz-Miller, T., Calderwood, K., Dragonetti, R. & Tsanos, A. (1997) Cognitive distortions in heavy gambling. *Journal of Gambling Studies* 13 (3):253–66. [aJMD]

Tooby, J. & Cosmides, L. (1990) The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology* 11:375–424. [SIH]

Trötschel, R. & Gollwitzer, P. M. (2007) Implementation intentions and the willful pursuit of prosocial goals in negotiations. *Journal of Experimental Social Psychology* 43:579–98. [MT]

Truth and Reconciliation Commission of Canada (2015) *Honouring the truth, reconciling for the future: Summary of the final report of the Truth and Reconciliation Commission of Canada*. Truth and Reconciliation Commission of Canada. [SIH]

Tuomela, R. (1995) *The importance of us: A philosophical study of basic social notions.* Stanford University Press. [BF]

Turiel, E. (2008) The development of children's orientations toward moral, social, and personal orders. *Human Development* 51:21–39. [SIH]

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D. & Wetherell, M. S. (1987) *Rediscovering the social group: A self-categorization theory*. Blackwell. [BF, SH]

Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–31. [JJC]

Uleman, J. K. (2010) *An introduction to Kant's moral philosophy*. Cambridge University Press. [JSU]

Uleman, J. S. (2015) Causes and causal attributions: Questions raised by Dave Hamilton and spontaneous trait inferences. In: *Social perception: From individuals to groups*, ed. S. J. Stroessner & J. W. Sherman, pp. 52–70. Psychology Press. [JSU]

Uleman, J. S., Rim, S., Saribay, S. A. & Kressel, L. M. (2012) Controversies, questions, and prospects for spontaneous social inferences. *Social and Personality Psychology Compass* 6:657–73. [rJMD, JSU]

Uz, I. (2015) The index of cultural tightness and looseness among 68 countries. *Journal of Cross-Cultural Psychology* 46(3):319–35. [BF]

Vaish, A. & Tomasello, M. (2014) The early ontogeny of human cooperation and morality. In: *Handbook of moral development,* 2nd edition, ed. M. Killen & J. G. Smetana, pp. 279–98. Psychology Press. [SIH]

van Baaren, R. B., Holland, R. W., Kawakami, K. & van Knippenberg, A. (2004) Mimicry and prosocial behavior. *Psychological Science* 15:71–74. [MT]

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J. & Reinero, D. A. (2016) Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113(23):6454–59. [aJMD]

Vargas, M. (2004) Responsibility and the aims of theory: Strawson and revisionism. *Pacific Philosophical Quarterly* 85(2):218–41. [aJMD]

Vargas, M. (2008) Moral influence, moral responsibility. In: *Essays on free will and moral responsibility*, ed. N. Trakakis & D. Cohen, pp. 90–122. Cambridge Scholars. [aJMD]

Vargas, M. (2013) *Building better beings: A theory of moral responsibility*. Oxford University Press. [aJMD, SM]

Vargas, M. R. (forthcoming-a) Reflectivism, skepticism, and values. *Social Theory and Practice*. [rJMD]

Vargas, M. R. (forthcoming-b) The social constitution of agency and responsibility: Oppression, politics, and moral ecology. In: *The social dimensions of responsibility*, ed. M. Oshana, K. Hutchinson & C. Mackenzie. Oxford University Press. [MRV]

Veenhoven, R. (1988) The utility of happiness. *Social Indicators Research* 20(4):333–54. [aJMD]

Velleman, J. D. (1989) *Practical reflection*. Princeton University Press. [aJMD]

Velleman, J. D. (2000) *The possibility of practical reason*. Oxford University Press. [aJMD]

Velleman, J. D. (2006) *Self to self: Selected essays*. Cambridge University Press. [aJMD]

Vohs, K. D. & Schooler, J. W. (2008) The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science* 19(1):49–54. doi:10.1111/j.1467-9280.2008.02045.x. [PMP]

von Hippel, W. & Trivers, R. (2011) The evolution and psychology of self-deception. *Behavioral and Brain Sciences* 34:1–16. [rJMD]

Voyer, B. V. & Franks, B. (2014) Toward a better understanding of self-construal theory: An agency view of the processes of self-construal. *Review of General Psychology* 18(2):101–14. [BF]

Wallace, R. J. (2003) Addiction as defect of the will: Some philosophical reflections. In: *Free will*, 2nd edition, ed. G. Watson, pp. 424–52. Oxford University Press. [aJMD]

Wallace, R. J. (2006) *Normativity and the will*. Oxford University Press. [aJMD]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press

Waller, B. N. (2011) *Against moral responsibility*. MIT Press. [TAM, PMP]

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K. & Ahn, H. N. (1997) A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "all must have prizes." *Psychological Bulletin* 122(3):203. [aJMD]

Wason, P. (1960) On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology* 12(3):129–40. [JJC]

Waters, F., Collerton, D., Jardri, R., Pins, D., Dudley, R., Blom, J. D., Mosimann, U. P., Eperjesi, F., Ford, S. & Larøi, F. (2014) Visual hallucinations in the psychosis spectrum and comparative information from neurodegenerative disorders and eye disease. *Schizophrenia Bulletin* 40:(Suppl 4):S233–45. [DC]

Watson, G. (1975) Free agency. *Journal of Philosophy* 72(8):205–220. [aJMD, JM, SM, MRV]

Watson, G. (1993) Responsibility and the limits of evil: Variations on a Strawsonian theme. In: *Perspectives on moral responsibility*, ed. J. M. Fischer & M. Ravizza, pp. 119–50. Cornell University Press. [aJMD]

Watson, G. (1996) Two faces of responsibility. *Philosophical Topics* 24(2):227–48. [aJMD]

Weaver, R. M. (1948) *Ideas have consequences.* University of Chicago Press. [TAM]

Webb, T. L. & Sheeran, P. (2002) Can implementation intentions help to overcome ego-depletion? *Journal of Experimental Social Psychology* 39:279–86. [MT]

Webb, T. L., Schweiger Gallo, I., Miles, E., Gollwitzer, P. M. & Sheeran, P. (2012) Effective regulation of affect: An action control perspective on emotion regulation. *European Review of Social Psychology* 23:143–86. [MT]

Webber, R., Rallings, C., Borisyuk, G. & Thrasher, M. (2014) Ballot order positional effects in British local elections, 1973–2011. *Parliamentary Affairs* 67(1):119–36. [aJMD]

Weeden, J. & Kurzban, R. (2014) *The hidden agenda of the political mind: How self-interest shapes our opinions and why we won't admit it*. Princeton University Press. [HV]

Wegner, D. M. (2002) *The illusion of conscious will*. MIT Press. [aJMD, PMP]

Wegner, D. M. (2005) Who is the controller of controlled processes? In: *The new unconscious*, ed. R. R. Hassin, J. S. Uleman & J. A. Bargh, pp. 19–36. Oxford University Press. [aJMD]

Weingarten, G. (2009) Fatal distraction. *The Washington Post*, March 8, 2009. Available at: https://www.washingtonpost.com/lifestyle/magazine/fatal-distraction-forgetting-a-child-in-thebackseat-of-a-car-is-a-horrifying-mistake-is-it-a-crime/2014/06/16/8ae0fe3a-f580-11e3-a3a5-42be35962a52_story.html?utm_term=.726005cd2fcb. [SM]

Westlund, A. C. (2003) Selflessness and responsibility for self: Is deference compatible with autonomy? *The Philosophical Review* 112(4):483–523. [rJMD]

Wieber, F., Gollwitzer, P. M. & Sheeran, P. (2014) Strategic regulation of mimicry effects by implementation intentions. *Journal of Experimental Social Psychology* 53:31–39. [MT]

Williams, B. (1993) *Shame and necessity*. University of California Press. [AP-C]

Wilson, R., Collerton, D., Freeston, M., Christodoulides, T. & Dudley, R. (2015) Is seeing believing? The process of change during cognitive–behavioural therapy for distressing visual hallucinations. *Clinical Psychology & Psychotherapy* 23:285–97. [DC]

Wilson, T. D. (2002) *Strangers to ourselves: Discovering the adaptive unconscious*. Belknap. [aJMD]

Wilson, T. D. & Schooler, J. W. (1991) Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology* 60:181–92. [NL]

Wolf, S. (1990) *Freedom within reason*. Oxford University Press. [SM]

Wong, D. (2006) *Natural moralities*. Oxford University Press. [MA]

Woodzicka, J. A. & LaFrance, M. (2001) Real versus imagined gender harassment. *Journal of Social Issues* 57:15–30. [DD]

Woolfolk, R. L. (1998) *The cure of souls: Science, values, and psychotherapy*. Jossey Bass. [aJMD]

Wootton, Anthony J. (1997) *Interaction and the development of mind*. Cambridge University Press. [JZ]

Wu, W. (2014) *Attention*. Routledge. [ZBW]

Yehuda, R., Hoge, C. W., McFarlane, A. C., Vermetten, E., Ruth A. Lanius, R. A., Nievergelt, C. M., Hobfoll, S. E., Koenen, K. C., Thomas C. Neylan, T. C. & Hyman, S. E. (2015) Post-traumatic stress disorder. *Nature Reviews* 1:1–21. [LN]

Zinken, J. (2016) *Requesting responsibility: The morality of grammar in Polish and English family interaction*. Oxford University Press. [rJMD, JZ]

Zinken, J. & Ogiermann, E. (2013) Responsibility and action: Invariants and diversity in requests for objects in British English and Polish interaction. *Research on Language & Social Interaction* 46(3):256–76. [rJMD]

Zuckerman, E. W. & Jost, J. T. (2001) What makes you think you're so popular? Self-evaluation maintenance and the subjective side of the "friendship paradox." *Social Psychology Quarterly* 64(3):207–23. [aJMD]

https://doi.org/10.1017/S0140525X16002016 Published online by Cambridge University Press