Response to Commentators

Neil Levy
Macquarie University and University of Oxford

Department of Philosophy
Macquarie University
NSW, 2109
Australia

Uehiro Centre for Practical Ethics
University of Oxford
Oxford, OX1 1PT
United Kingdom

neil.levy@philosophy.ox.ac.uk
ORCID: 0000-0002-5679-1986

Abstract: This paper replies to the contributors to a symposium on the book *Bad Beliefs*. It groups the criticisms and concerns of the contributors under the headings 'Gaps and Holes', 'Rationality', 'Epistemic Virtue', 'Agency and Control' and 'Nudges'. It defends the view that bad belief formation and maintenance is very importantly rational, though it also acknowledges gaps, limitations and unanswered questions.

Keywords: social epistemology; rationality; evidence; cultural evolution.

In *Bad Beliefs*, I defended a deeply social conception of rational cognition: good thinking is distributed across agents and over time. This is true at the level of cultures, I believe, but also at the level of disciplines and subdisciplines: they're truth conducive, when they are, because they're conversations. I am very grateful that such a wonderful set of conversationalists have taken up the opportunity to talk with me. Consistent with my view that we're extremely limited on our own, they've shown up problems, limitations, and faults in my work. Philosophy being philosophy, they've also raised concerns that reflect big picture disagreements between my views and those of the commentators. In particular, a number of commentators reject the central claim of the book: that bad beliefs arise, for the vast majority of those who hold them, through rational processes. Rather than address the concerns of each commentator separately, I've gathered them under headings, to bring out commonalities as well as to highlight differences.

Gaps and Holes

I'll begin by addressing the gaps that commentators pointed to in the book. I aimed to write a short book, and (as a defender of distributed cognition) didn't see it as imperative to be comprehensive. Nevertheless, commentators identified holes that might be better filled.

Catarina Dutilh Novaes' biggest concern about *Bad Beliefs* is what she sees as its neglect of the political dimension of human life. It's true that there's little direct engagement with that dimension; nevertheless, I see the book as addressed to it. I take the political and social to be more or less co-extensive: talk of the political is a way of emphasising the stakes and the role of values and of power in the social. In focusing on how our informational environment is structured, I take myself to be concerned with the political. I think Dutilh Novaes' deeper concern might be with what she sees as a naivety in my (admittedly very sketchy) recommendations for stewardship of the epistemic environment. We have no right to optimism about my suggestions in the face of strategic resistance from epistemic polluters, she suggests. I completely agree. Dutilh Novaes accuses me of an "overly rosy" account but actually I'm far from optimistic. While I'm sure others can design better strategies for stewardship than I've been able to, I'm deeply pessimistic that they'll succeed in making much of a dent on the problems we confront.

Dutilh Novaes also points to a different gap: a lack of engagement with existing literature on the concept of higher-order evidence. As she notes, my usage is non-standard, and requires a defence. She offers one on my behalf: my usage might be justified on the grounds that since disagreement counts as higher-order evidence, so should agreement. I think this defence only justifies some of the ways in which I use the term. Disagreement is evidence about evidence; in cases like 'Restaurant check' (Christensen 2007), peer disagreement is evidence that the agent has processed her first-order evidence badly. In other cases, disagreement might indicate not that that the agent has processed her evidence badly but that she's missed evidence. Conversely, agreement may be evidence we've processed our evidence well, or that we haven't missed relevant evidence. But social cues and even testimony only sometimes provide evidence of these sorts.

Sometimes social cues are provided and responded to without the agent having any (first-order) evidence at all. Similarly, agents may be recipients of testimony that *p* without having any (other) evidence whether *p*. *Higher* is a relational term; evidence counts as 'higher' in relation to lower order evidence. In these cases, the agent lacks lower-order evidence, so there's some reason to resist describing these cases as instances of higher-order evidence. But on my deeply social model, I suggest that we can appropriately describe these cases in this way, because they involve the provision of evidence about *someone else's* evidence. When I recommend a restaurant, I provide evidence that *my* evidence supports believing that it's good: the agent who is the recipient of such testimony is therefore in receipt of higher-order evidence.

That brings us to another gap: an absence of a discussion of attention, despite my frequent invocation of salience as higher-order evidence. Dutilh Novaes shows how salience may dovetail illuminatingly with the three-tiered model of epistemic exchange she sets out. Altering the *salience structure* of information (a term Dutilh Novaes borrows from Munton (2021)) channels attention: it provides higher-order evidence *that this is important*. The salience structure might be an extra-individual realization of prejudice (Dutilh Novaes' reflections on attention interestingly complements recent work by Georgi Gardiner (2022), focusing on how allocations of attention can be vicious at the individual level). As Dutilh Novaes emphasizes, attention to the *harms* of salience structures complicates my suggestion that salience functions as a recommendation. We may make something salient to criticize it, as well as to recommend

it. She's right: I should have attended to attention, and in particular to its harms. It's important to distinguish the different valences of recommendations: that's indeed a gap in *Bad Beliefs.*

Rie Iizuka and Chie Kobayashi point to a different gap: they argue that I am not sufficiently cognizant of epistemic injustice, especially of what they (following Grasswick, 2017) call *participatory epistemic injustice* (the epistemic injustice consisting in denying members of specific groups the opportunity to participate in cooperative activities as epistemic agents). They're right: I didn't address epistemic injustice in *Bad Beliefs*: neither Fricker's original testimonial and hermeneutical injustices, nor Grasswick's participatory injustice. Moreover, they're right that these injustices – especially testimonial and participatory injustice – are not harms I can safely set aside: since one of my aims is restoring trust in science, and these injustices not only lower the quality of science but also *rationally* reduce trust in it, addressing these injustices should have figured among the proposals for restoring trust I outlined in chapter 5. Members of marginalized communities rationally distrust institutions and professionals who don't take their testimony seriously, who don't address their concerns and who may even exploit their epistemic capital for their own ends.

I'm not too embarrassed by this gap: I should have noted the importance of the topic, but I'm not the right person to address it in detail. I hoped *Bad Beliefs* would participate in a conversation that began long before I wrote it and will continue long after it is forgotten; I hoped to direct that conversation to neglected perspectives and views. If my hope is fulfilled, then other people will engage with it, note its limitations and – hopefully – go beyond them, as Iizuka and Kobayashi have done.

Rationality

My sense is that these gaps can be remedied without major disruption elsewhere in the overall picture. A more worrying problem concerns the *rationality* of seeing cues as recommendations. This is an issue that worries both Dutilh Novaes and Dan Williams. Williams takes *Bad Beliefs* to be even more confused than I fear in my darker moments. He thinks that I appeal to processes I take to be "unsophisticated and credulous" to underwrite the rationality of belief formation (even worse for me, Williams thinks that not only do I take people to be "dim-witted and docile", I also – entirely inconsistently – present bad arguments for why they are intelligent), None of this sounds very promising, and Williams is right to reject the package. But of course I don't take myself to appeal to unsophisticated and credulous processes or dim-witted people (and of course I don't take myself to be inconsistent in presenting arguments for why they're intelligent, bad arguments or not).

Williams argues that social learning as depicted by the West Coast school of cultural evolution is committed to credulous mechanisms. I deny it is (see Boudry, 2018; Buskell, 2016; Levy and Alfano, 2019). It might be best simply to set aside this issue, though and let the debate between different schools in cultural evolution drop out. Whether these particular mechanisms are intelligent or not, I *do* appeal to intelligent mechanisms – indeed, I cite the rival, Parisian, school extensively in discussing the details of those mechanisms. Everything Williams says about the intelligence of these mechanisms I gladly echo ("people evaluate the plausibility of group opinions, the degree to which they are formed independently, the competence and benevolence of group members, the size of the group, and the personal costs of error"). The

claim that the rationality of social learning is underwritten by rational processes obviously looks in much better shape than the view that Williams attributes to me, according to which I appeal to unintelligent mechanisms to ground the intelligence of people.[1]

Dutilh Novaes' worries about the rationality of relying on higher-order evidence turn on concerns about the possibility of a lack of independence between testifiers/recommenders and the possibility that the environment has been manipulated. I believe that a lack of independence is less of a problem than many people think. What Goldman (2001) calls "non-discriminating reflectors" are in fact vanishingly rare: because people assess information for plausibility as well as by assessing the competence and benevolence of informants (Harris 2012; Sperber et al. 2010), each agent adds value to testimonial networks. These points are relevant to the possibility of misinformation too. People are taken in less often than we tend to think (Mercier 2020): we're reasonably good at detecting deception. It's when we take ourselves to have no other relevant information and we're not motivated to check that we're most liable to be deceived. In light of this, canny deceivers turn their attention to people's motivations: to convince people of misinformation, it's first necessary (or at any rate immensely helpful) to convince them that it doesn't matter much.

These reflections are directly relevant to responding to another of Williams' concerns. He points out that informational conformity depends on putting weight on the opinions of non-experts. But non-experts' views are neither independently formed nor well informed. In my view, testimonial networks are rationally taken to add epistemic value to the information that propagates through them. It's rational to defer to networks that consist of laypeople when there's good reason to think they're appropriately plugged into experts, and that's a condition that the kind of networks we're concerned with here satisfy. The ordinary conservative's dismissal of climate change is multiply buttressed: by conformity with her peers (whose behavior and assertions are best explained by attributing to them a matching belief) and through testimony from news sources and politicians. These news sources and politicians, in turn, she has good reason to believe, receive testimony from genuine experts. She may rationally conform to the beliefs of those around her, secure in the knowledge that these networks transmit and test expert testimony. When a belief that reverberates around this network is near-universally held with high confidence, her deference is rational. It is also intelligent: she remains sensitive to signs of dissent from those who possess expertise and who are trustworthy (by her lights). Under a variety of conditions, her confidence will fall: for example, when the proportion of believers in the network shows signs of dropping (it's worth noting, by the way, that this model may realize the kind of responsible 'doing one's own research' advocated by Murphy-Hollies and Caporuscio).

As I argued in *Bad Beliefs*, theory choice is a comparative affair, and Williams argues in favor of a more traditional and less rational explanation of bad beliefs. On his *incentives* account, bad beliefs arise when the costs of error are low and the incentives for being biased are high (note that the *short-term* costs of error must be low on my model, too. That's because short-term costs are an error signal, and agents intelligently combine such signals in coming to their beliefs). He gives several reasons to prefer his account to mine. First, it explains what he takes to be the domain-specificity of irrationality: why it is confined to "domains such as politics". Obviously, I don't accept that political thought is irrational. In any case, I think we see the same sort of deference and (rationally) superficial thinking everywhere: it's just that we tend

to pay more attention to politicised areas, so we notice conflict with the evidence more strongly in that domain. The illusion of explanatory depth, driven by rational outsourcing, shows up in politics (Sloman and Vives 2022), but is utterly pervasive (Keil and Wilson 2000; C. M. Mills and Keil 2004; Rabb, Fernbach, and Sloman 2019).

Second, Williams argues that my view has uncomfortable implications: I would be forced to accept that, say, white supremacist ideology is epistemically rational for many people. If the epistemic authorities in a society testify that such an ideology is rational, and most people conform to it, then on my view a layperson should take it to be justified. Worse, I seem committed to thinking that dissenters and reformers are *irrational*. Williams cites Mary Wollstonecraft, who rejected the elite consensus that women were cognitively inferior. These, I'm afraid, are bullets I'm willing to bite.

Extreme ideologies *are* often subjectively rational. The ordinary person may have a great deal of (higher-order) evidence in their favor, and typically little evidence against them. She may, for instance, have little contact with members of oppressed groups, except in conditions where they're required to perform servility. Moreover, oppression conspires to produce evidence that justifies it. It often deprives the oppressed of the opportunities to develop sophisticated cognitive tools and even undermines moral virtue and self-control. It's harder to be morally virtuous when life is a struggle: selfishness might be a survival strategy. Cram the poor into slums and it won't be hard to find evidence that they're grasping and dirty.

Given these facts, rational dissent is harder to come by than we often naively imagine. As Mills (2017) suggests, social conditions play a big role in explaining whether the oppressed develop the conceptual tools for making sense of their oppression. He argues that the segregation of black people in the United States allowed them to develop such tools: they could exchange, test and refine ideas, and probe conditions together with other people they had good reason to trust. All of this was harder for white women, because patriarchal ideology kept them separated from one another. In general, we should expect the beneficiaries of oppression to rationally accept it, not (just) because they're motivated to, but because they lack access to countervailing voices that they have good reason to trust. Those who are oppressed are in a better position to see what's wrong with their oppression, because it's much harder to convince them that those who promulgate such ideologies have their interests genuinely at heart. But their dissent may often be piecemeal and partial. That's not to say that Wollstonecraft was irrational: rather, it's to say that her being able to develop a sophisticated critique of the oppression of women was dependent on her occupying a rare epistemic situation in which orthodox opinion rationally carried less weight for her than it did for most women.

Stephen Gadsby presses closely related concerns: he cites a range of empirical evidence that, he claims, shows that many of the mechanisms involved are *a*rational, not rational. He cites the *illusory truth effect* – the increase in credence assigned to a proposition subsequent on its repetition – and Daniel Gilbert's evidence in support of the Spinozan theory of belief formation. On the Spinozan theory, belief formation is automatic: we always believe every proposition we entertain, and only subsequently reject some when we have the time and resources to subject them to assessment. In one experiment, participants were given nonsense (ostensibly Hopi) words, together with a definition, followed by the further information that

the definition was true or false (this was study 1; studies 2 and 3 were conceptual replications). Interruption of processing had no effect on participants' assessment of true statements but increased their rates of misclassification of false statements as true (Gilbert, Krull, and Malone 1990). Along similar lines, Wegner et al. (1985) provide evidence for belief perseverance in the face of discounting evidence. Even after being told that feedback was (or in one condition, would be) predetermined and did not reflect actual performance, participants continued to give it weight in assessing how well the agent did and would do at the task.

The first thing to say about the evidence from Gilbert and Wegner is that I have low confidence in it. It stems from the bad old days of social psychology, when preregistration was non-existent, sample sizes were small and questionable research practices were rife (Gilbert is a vocal opponent of reforms aimed at eradicating these practices; Gilbert et al. 2016). These concerns apply generally, but given the sample sizes in these experiments, I think they apply to this work specifically. It's noteworthy, too, that later research has cast doubt on the Spinozan theory. Automatic acceptance, if it occurs at all, seems to occur only when the person has no relevant background information and no reason to care about the truth of the statement. When it matters, they devote cognitive resources to processing it, and do so rationally (see Mercier, 2017 for discussion). I am far from convinced that it's arational not to devote cognitive resources to irrelevant nonsense.

I put significant more weight on the illusory truth effect, largely because it has been replicated by researchers who seem reliable. But the effect also looks pretty rational to me. Numbers count when it comes to belief formation: the more sources of a claim there are, the more likely it is to be true. Familiarity is a proxy for numbers: unsurprisingly we respond to it. It's an interesting question whether undercutting evidence, as in the Wegner debrief (or Gadsby's proposed randomization of defaults), should reduce the force of the evidence to zero (I'm confident it would reduce the force of the evidence, as Wegner himself notes). That depends, inter alia, on how confident participants are in the relative credibility of the evidence. A good test will use participant-relevant information: here I record my prediction that participants will be discerning.

I'm not committed to the view that cognitive mechanisms are always rational. Perhaps – *perhaps* – some are arational as Gadsby suggests (perhaps advocacy effects are an example, though it's not obvious that it's arational to be disposed to think favorably of someone you are preparing to advocate for – perhaps it's rational, or perhaps it simply fails to fall under epistemic norms at all).[2] My view is that most such mechanisms are responsive to evidence and that as a consequence most bad beliefs arise through rational processes. Where Gadsby sees arationality, I see rationality. Where he sees *ir*rationality – as in his examples of the ways in which presidents Bush and Johnson structured their informational environment – I see arationality *at worst*. All we need suppose is that each president believed (surely truly) that listening to people who held a view contrary to their own would reduce his confidence in that belief. As I argued in the book, dogmatism is sometimes rational. I'm dogmatic with regard to climate change sceptics: I don't listen to what they have to say. Why should I deliberately bring it about that my confidence in an extremely well-justified belief is reduced?

Epistemic virtue

Williams claims that "we know the kinds of characteristics that are conducive to forming accurate beliefs:" the epistemic virtues such as humility, curiosity, self-criticism, acknowledgement of uncertainty, and most of all a willingness to treat beliefs as hypotheses to be tested, rather than possessions to be defended. Williams cites Tetlock (Tetlock and Gardner 2016) in this context, but there's a great deal of other evidence apparently showing that the virtues conduce to better thinking. For example, Meyer et al. (2021a) found that epistemic vice correlates with the acceptance of misinformation about Covid, and the same researchers found it predicted conspiratorial ideation (Meyer et al. 2021b). I am unconvinced either set of findings present an insuperable challenge for my view.

It's worth emphasising that Tetlock's research focuses on what Kahneman and Klein (2009) call low-validity domains: domains in which feedback is not rapid enough or reliable enough for expert judgment to be well calibrated. It's because, say, political forecasting is a low-validity domain (because political outcomes are influenced by so many disparate factors and because it's difficult to measure the causal contributions of each) that expert judgment in this domain is no more accurate than non-expert judgment. Superforecasters outperform the experts, but they don't do very well absolutely: this just isn't a domain in which it's possible to be reliable. These facts make me suspicious that these domains are good models for assessing the contribution of epistemic virtues – it's supposed to be relatively easy to see that the virtues conduce to good thinking.

I'm also sceptical that superforecasters' greater success is due to the *virtues*: domain-general character traits. In large part at least, they seem due to teachable heuristics, like considering base rates and comparison classes. In combination, these concerns leave me relatively comfortable with the challenge from Tetlock: this work doesn't show that epistemic virtues lead to better performance in the domains I'm concerned with. In high-validity domains, the distribution of cognition across networks of researchers with divergent epistemic interests and dispositions is powerful enough to render the effects of virtues insignificant in comparison. Non-experts don't need the virtues because they should and do defer to experts with regard to these domains. Outside these domains, no one has any right to great confidence.

Meyer et al. might seem the bigger challenge: the questions they're concerned with either stem from high-validity domains or are easy questions within low-validity domains. Their evidence might therefore be taken to show that epistemic virtue correlates with better deference. That's a challenge for me, though not quite the one Williams has in mind. I'm sceptical, though, that Meyer et al. actually measure epistemic virtue. The scale they use is transparent to responders: it's quite apparent what's being measured. I suspect responses probe being disposed to give the answers that (as the respondents well understand) are expected of them. What drives variance in responses is more likely a disposition to troll (Lopez and Hillygus 2018) or to engage in expressive responding (Hannon 2021; Levy and Ross 2021) than epistemic vice per se.

Kathleen Murphy-Hollies and Chiara Caporuscio also suggest I've neglected the epistemic virtues (this is an element of their broader claim that I neglect the agent more generally). I do think there's a role for the virtues in explaining belief formation, but it's a much smaller role than virtue epistemologists think. Of course, it's true that belief formation depends on internal

processing as well as the environment, and we can perspicuously describe aspects of that processing in the language of the virtues. But I don't think we get great bang for our buck by focusing on the virtues. First, I think there's relatively little variance in most of the virtues; more importantly, I doubt that the variance we do see explains epistemic success and failure to any great extent. Second, I am sceptical that attempts to inculcate the virtues, thought of as individual dispositions, are likely to make much difference to the variance. If we want to improve people's *internal* functioning, I suspect, we'd do better to focus on their environment, and especially on how rewards are distributed in the environment (see Karabegovic and Mercier 2023 for discussion of how environments can encourage intellectual humility in particular). Reward conscientiousness and I think you'll get better results than by focusing directly on the virtues. The only virtue (or rather vice) term I suspect plays a significant explanatory role is intellectual arrogance, but variance in arrogance is itself explained by social status and the costs of displays.

Agency and control

That brings us directly to Murphy-Hollies and Caporuscio's broader concern, that I leave the agent out of the picture. As they argue, belief formation is never simply a function of the environment agents find themselves in, but also a function of the agent herself. If we plonk a left-leaning scientist in a conservative environment in which everyone believes climate change is a hoax (to use their own example), they're not suddenly going to reject their former view. They argue that this shows that beliefs aren't 'shallow', in the sense I used the term in *Bad Beliefs*; i.e., liable to sudden shifts in response to external cues. Putting the agent back into the environment is necessary, they argue, in order to understand irrationality, the role of epistemic virtues and vices, and agents' responsibility for their beliefs.

As I understand their criticisms, if we follow my lead and focus on the environment to the (near) exclusion of the agent, we'll be unable to see how individuals can make certain sorts of really significant epistemic contributions. For example, they argue that the difference between the agent whose inherited epistemic resources are reliable and the agent who lacks such resources must come down to luck on my view. Without individual-level understanding of mechanisms, we cannot know whether we're in the good case or the bad. We need "good old fashioned individual rationality" to distinguish the cases, but I seem to deny the agent such resources.

They're right: that's an implication of my view, and it's a bullet I'm willing to bite. On my view, being in the good case is indeed good luck. There is of course a story to tell about why it is the good case – about how we're the beneficiaries of knowledge-production and transmission processes that are distributed across agents and over time – but were we in the bad case, we'd think (wrongly but rationally) that we had such a story to tell about our own resources, or a rival story that was justified for us. There are, as Murphy-Hollies and Caporuscio rightly point out, deep questions here about how individuals may or can contribute to cultural knowledge production. I have no good answers to these questions; so far as I can tell, no one else does either. On my view, for what it's worth, individual innovation must be very local: agents should innovate only in the very specific area of their special expertise. We should avoid epistemic trespassing (Ballantyne 2019), and we should understand 'trespassing' very broadly. Since I argue that even the expert is dependent on

testimony even within the sphere of their own expertise, the expert, too, is a beneficiary or victim of luck: whether her innovations improve an already reliable epistemic mechanism or merely add an epicycle to a model that fails to track reality is something she may not be able reliably to discern.

It is absolutely true, as Murphy-Hollies and Caporuscio point out, that our internal dispositions don't simply change as a function of changes in our environment; nevertheless, they are very significantly a product of our past environment, and they can be expected to change in response to changes in the current environment (if the left-leaning scientist establishes friendly relations with their conservative neighbors, their beliefs will likely shift gradually, since the behavior of those we respect provides us with evidence). Our beliefs are shallow not in that they will shift in response to any and all changes in epistemic cues, but they will respond rapidly and effortlessly if cues we regard as authoritative alter.

Gadsby shares with Murphy-Hollies and Caporuscio the worry that I underplay the amount of control individuals have over their epistemic environments. While we do exercise control, there's a regress in the offing. Control has epistemic conditions (Levy 2011): we exercise control only in the light of attitudes that rationalize actions, including epistemic actions. That entails that our credences, together with environmental input, play a determinative role in belief formation. We don't exercise control in a way that underwrites a reasonable expectation that we avoid epistemic risks or exit our testimonial network; not under typical circumstances. Nor does our control make us *active* in belief formation; not in the sense that Murphy-Hollies and Caporuscio have in mind. It's not far from the truth, in my view, to say that beliefs *happen* to people: our only activities are the direction of attention and the gathering of evidence, and these are actions which are themselves caused by our beliefs. It follows, as Murphy-Hollies and Caporuscio point out, that it is hard to make sense of agents' epistemic and moral responsibility. I see that as a benefit, not a cost, of my view.

Nudges

The final issue I'll focus on concerns my account of nudges. Rie Iizuka and Chie Kobayashi argue that this account neglects some of the ways in which nudges might threaten our autonomy. They argue that for nudges to be compatible with autonomy, they must not merely be non-manipulative (as I claim they are); they must also be transparent. If a nudge is not transparent, they argue, it's difficult to resist it, and the nudged person cannot judge whether the nudge promotes their own conception of the good. My claim is that nudges simply provide recommendations: there's no *special* problem that nudges raise. It is often a good thing if we're transparent with our recommendations in the sense Iizuka and Kobayashi mean: that is, if we tell people that we're recommending something because it promotes a particular end, but that's not a problem that nudges raise specifically. Moreover, it's often not needed, because the context makes it clear. Selecting a default option on a retirement savings plan is understood by people as promoting the right balance between income now and in retirement; there's no need to communicate any additional information.

Iizuka and Kobayashi raise another concern about my treatment of nudges. They're right: I neglected the important question of who should design nudges. People may have legitimate concerns that nudges are in the interests of particular sections of the population and not the

groups to which they belong, and these concerns might (rationally) undermine trust in the nudges and the institutions that promote them. We need more diverse voices in the design of nudges so that the interests of the marginalized are not ignored. Restoring trust in science and in institutions, and making them work in the interests of the entire population, are goals that reinforce one another: better design, more diversity, more inclusion and more consultation will support both.

Conclusion

A central message of *Bad Beliefs* is that many minds are almost always better than one: exchange, criticism, thinking together and even against one another is how knowledge advances. For me, at any rate, these exchanges have exemplified the epistemic benefits of distributed cognition: I've learned a lot from the commentators. I've also been enabled to better understood my own views, as well as its limitations. Alas, the commentators haven't succeeded in uprooting many of my bad beliefs: I continue to believe that belief formation is largely rational, even when it's out of touch with reality. No doubt I'm less in touch with reality than I think; whether I'm rational nevertheless is something for others to judge.[3]

NOTES

[1] The intelligence of these processes is orthogonal to another question on which Williams takes me to task: the role of intelligence in the elaboration of cultural traditions (this is a concern shared by Murphy-Hollies and Caporuscio). It's true I have no solution to the problems he mentions, but neither does anyone else. An appeal to intelligence is no solution at all. Of course, all sides acknowledge that innovation is usually intelligent, but what we're trying to explain is innovation that builds upon practices that are opaque to those who use them. People can build intelligently on something they don't understand, but they have a very limited capacity to intervene intelligently in its workings.

[2] Gadsby characterizes processes that don't fall under epistemic norms as arational. I'm not sure this is a useful way of characterizing them. We don't want to say that photons and ice crystals are arational. I think it's more perspicuous to reserve "arational" for processes and agents that do fall under epistemic norms but are neither irrational nor rational.

[3] I'm grateful to the commentators for their thought provoking contributions to this symposium, and Lisa Bortolotti for her sterling editorial work. I am also grateful to the John Templeton Foundation (grant #62631) and the Arts and Humanities Research Council (AH/W005077/1) for their support.

References

Ballantyne, Nathan. 2019. *Knowing Our Limits.* Oxford: Oxford University Press. https://www.amazon.co.uk/Knowing-Our-Limits-Nathan-Ballantyne/dp/019084728X/ref=sr_1_1?keywords=nathan+ballantyne&qid=1569462767&s=books&sr=1-1.

Boudry, Maarten. 2018. 'Replicate after Reading: On the Extraction and Evocation of Cultural Information'. *Biology & Philosophy* 33 (3): 27. https://doi.org/10.1007/s10539-018-9637-z.

Buskell, Andrew. 2016. 'Cultural Longevity: Morin on Cultural Lineages'. *Biology & Philosophy* 31 (3): 435–46. https://doi.org/10.1007/s10539-015-9506-y.

Christensen, David. 2007. 'Epistemology of Disagreement: The Good News'. *The Philosophical Review* 116 (2): 187–217. https://doi.org/10.1215/00318108-2006-035.

Gardiner, Georgi. 2022. 'Attunement: On the Cognitive Virtues of Attention'. In *Social Virtue Epistemology*, 48–72. Routledge.

Gilbert, Daniel T., Gary King, Stephen Pettigrew, and Timothy D. Wilson. 2016. 'Comment on "Estimating the Reproducibility of Psychological Science"'. *Science* 351 (6277): 1037–1037. https://doi.org/10.1126/science.aad7243.

Gilbert, Daniel T., Douglas S. Krull, and Patrick S. Malone. 1990. 'Unbelieving the Unbelievable: Some Problems in the Rejection of False Information.' *Journal of Personality and Social Psychology* 59 (4): 601.

Grasswick, Heidi. 2017. 'Epistemic Injustice in Science'. In *The Routledge Handbook of Epistemic Injustice*. Routledge.

Hannon, Michael. 2021. 'Disagreement or Badmouthing? The Role of Expressive Discourse in Politics'. In *Political Epistemology*, edited by Elizabeth Edenberg and Michael Hannon. Oxford University Press. https://www.academia.edu/40013480/Political_Disagreement_or_Partisan_Cheerleading_The_Role_of_Expressive_Discourse_in_Politics.

Harris, Paul. 2012. *Trusting What You're Told.* Harvard University Press.

Kahneman, Daniel, and Gary Klein. 2009. 'Conditions for Intuitive Expertise: A Failure to Disagree'. *American Psychologist* 64 (6): 515–26. https://doi.org/10.1037/a0016755.

Karabegovic, Mia, and Hugo Mercier. 2023. 'The Reputational Benefits of Intellectual Humility'. *Review of Philosophy and Psychology*, February. https://doi.org/10.1007/s13164-023-00679-9.

Keil, Frank C., and Robert A. Wilson. 2000. 'The Shadows and Shallows of Explanation'. In *Minds and Machines*, edited by Frank C. Keil and Robert A. Wilson, 137–59. MIT Press.

Levy, Neil. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press. https://www.amazon.com/Hard-Luck-Undermines-Moral-Responsibility-ebook/dp/B006SVNNZO/ref=sr_1_1?keywords=hard+luck+levy&qid=1575370264&s=books&sr=1-1.

Levy, Neil, and Mark Alfano. 2019. 'Knowledge From Vice: Deeply Social Epistemology'. *Mind*. https://doi.org/10.1093/mind/fzz017.

Levy, Neil, and Robert M. Ross. 2021. 'The Cognitive Science of Fake News'. In *The Routledge Handbook of Political Epistemology*, edited by Michael Hannon and Jeroen de Ridder, 181–91. Routledge.

Lopez, Jesse, and D. Sunshine Hillygus. 2018. 'Why So Serious?: Survey Trolls and Misinformation'. SSRN Scholarly Paper ID 3131087. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3131087.

Mercier, Hugo. 2017. 'How Gullible Are We? A Review of the Evidence from Psychology and Social Science'. *Review of General Psychology* 21. https://journals.sagepub.com/doi/10.1037/gpr0000111.

———. 2020. *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton: Princeton University Press.

Meyer, Marco, Mark Alfano, and Boudewijn de Bruin. 2021. 'The Development and Validation of the Epistemic Vice Scale'. *Review of Philosophy and Psychology*, June. https://doi.org/10.1007/s13164-021-00562-5.

Meyer, Marco, Mark Alfano, and Boudewijn de Bruin. 2021. 'Epistemic Vice Predicts Acceptance of Covid-19 Misinformation'. *Episteme*, July, 1–22. https://doi.org/10.1017/epi.2021.18.

Mills, Candice M, and Frank C Keil. 2004. 'Knowing the Limits of One's Understanding: The Development of an Awareness of an Illusion of Explanatory Depth'. *Journal of Experimental Child Psychology* 87 (1): 1–32. https://doi.org/10.1016/j.jecp.2003.09.003.

Mills, Charles W. 2017. 'Ideology'. In *The Routledge Handbook of Epistemic Injustice*, edited by Ian James Kidd, Jose Medina, and Gaile Pohlhaus, 100–111. Routledge.

Munton, Jessie. 2021. 'Prejudice as the Misattribution of Salience☆'. *Analytic Philosophy* n/a (n/a). https://doi.org/10.1111/phib.12250.

Rabb, Nathaniel, Philip M. Fernbach, and Steven A. Sloman. 2019. 'Individual Representation in a Community of Knowledge'. *Trends in Cognitive Sciences* 23 (10): 891–902. https://doi.org/10.1016/j.tics.2019.07.011.

Sloman, Steven A., and Marc-Lluis Vives. 2022. 'Is Political Extremism Supported by an Illusion of Understanding?' *Cognition* 225 (August): 105146. https://doi.org/10.1016/j.cognition.2022.105146.

Sperber, Dan, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. 'Epistemic Vigilance'. *Mind & Language* 25 (4): 359–93. https://doi.org/10.1111/j.1468-0017.2010.01394.x.

Tetlock, Philip E., and Dan Gardner. 2016. *Superforecasting: The Art and Science of Prediction*. Random House.

Wegner, Daniel M., Gary F. Coulton, and Richard Wenzlaff. 1985. 'The Transparency of Denial. Briefing in the Debriefing Paradigm'. *Journal of Personality and Social Psychology* 49 (2): 338–46.