Improving knowledge acquisition and dissemination through technological interventions on cognitive biases

Stammers, Sophie

DOI:
10.1111/edth.12340

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Stammers, S 2019, 'Improving knowledge acquisition and dissemination through technological interventions on cognitive biases', *Educational Theory*, vol. 68, no. 6, pp. 675-692. https://doi.org/10.1111/edth.12340

EDUCATIONAL THEORY

IMPROVING KNOWLEDGE ACQUISITION AND DISSEMINATION THROUGH TECHNOLOGICAL INTERVENTIONS ON COGNITIVE BIASES

Sophie Stammers

Department of Philosophy
University of Birmingham

Abstract. Much of the philosophical debate regarding technological epistemic enhancement concerns interventions on cognitive capacities that are already performing well in order that they perform even better. However, several decades of research in cognitive science demonstrates that humans harbor systematic cognitive biases that can produce ill-grounded, distorted, or otherwise epistemically faulty cognitions. Such cognitions can be stubborn and so may be good candidates for reduction or elimination through technological intervention. In this article, Sophie Stammers demonstrates that we can take two approaches in such an endeavor: interventions that halt and redirect the general processes that can generate biased cognitions; and targeted interventions that aim to extinguish or overwrite individual biased cognitions. Stammers argues that, because the general processes that produce biased cognitions also regularly produce accurate cognitions, an intervention that halts these processes altogether will have the result that the person undergoing the intervention will find it difficult to form new beliefs and acquire further knowledge, an epistemically undesirable outcome. Targeted interventions are therefore preferable because they enable us to eliminate some biases, while leaving otherwise useful processes (such as heuristics) intact. She concludes by demonstrating that our epistemic goals will be best achieved when targeted technological interventions are supplemented by attention to relevant structural features.

Introduction

Discussion of both the possibility and desirability of using technological developments to enhance human performance in a variety of domains occupies theorists across a wide range of disciplines. A subset of these technologies aims to enhance the process of knowledge acquisition ("epistemic enhancement," to use John Danaher's term), raising important questions for both educational theorists and practitioners.[1] Much of the philosophical debate regarding technological enhancement in the epistemic realm concerns interventions on cognitive capacities that are already performing well in order that they perform even better.[2] These discussions are about *extending* the boundaries of human performance. However, several decades of research in cognitive science demonstrates that humans harbor systematic cognitive biases that can produce ill-grounded, distorted, or otherwise epistemically faulty cognitions. These biases have a range of deleterious effects on multiple aspects of both teaching and learning. While philosophers have discussed the possibility and permissibility of technological epistemic enhancement of cognitive capacities that are already performing well, whether technological interventions might be used to mitigate cognitive biases is underexplored.[3]

1. John Danaher, "On the Need for Epistemic Enhancement: Democratic Legitimacy and the Enhancement Project," *Law, Innovation, and Technology* 5, no. 1 (2013): 85–112.

2. See, for example, Nick Bostrom and Anders Sandberg, "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges," *Science and Engineering Ethics* 15, no. 3 (2009): 311–341.

3. Discussion of technological interventions that play an ameliorative role has tended to focus on therapeutic applications — the use of propanol to reduce anxiety, for example — that may have indirect

In this article, I investigate the use of future technological interventions to mitigate the effects of cognitive bias in education. I first consider current research that may pave the way for future technological cognitive interventions (section 1). I then introduce cognitive bias (section 2) and focus on two examples, confirmation bias (2.1) and social bias (2.2), and justify why these are of particular concern in educational settings. In section 3, I demonstrate that we can take two approaches to addressing such biases through technological cognitive interventions: (a) interventions that halt and redirect the *processes* that generate biased cognitions; and (b) *targeted* interventions that aim to extinguish or overwrite individual biased cognitions. I argue that, because the processes that produce biased cognitions also regularly produce accurate cognitions, an intervention that halts these processes altogether will have the result that the person undergoing the intervention will find it difficult to form new beliefs and acquire further knowledge, an epistemically undesirable outcome. Targeted interventions are therefore preferable because they enable us to eliminate some biases while, at the same time, leaving often useful heuristic processes intact. However, such interventions could have the result of obscuring information regarding social and historical injustices that, in part, explain differential achievements — information that is of vital import to effective teaching, for example, because it contextualizes and justifies the provision of support. I suggest that interventions that enable the selective erasure of social biases should invite the user to consider the wider social and historical context of these biases and should demonstrate how this facilitates progress in research and also enhances the dissemination of knowledge in teaching and learning more generally.

1. Kinds of Technological Epistemic Enhancement

Interventions with numerous different kinds of technologies may result in enhanced cognition that facilitates knowledge acquisition. The locus of interest in this section is not necessarily with technology that produces a *distinctive* form of technologically enhanced cognition. My concern is with identifying technological interventions that are reasonably novel, meaning that they have not been

epistemic benefits by making a person more comfortable in a learning environment; or the use of central nervous system stimulants to improve attention in learning environments for people with ADHD (Lawrence H. Diller, "The Run on Ritalin: Attention Deficit Disorder and Stimulant Treatment in the 1990s," *Hastings Center Report* 25, no. 2 [1996]: 12–18.) There is less discussion of whether individuals could, and should, employ technological interventions to improve their systematically faulty cognition, regardless of whether they are neurotypical or not. One exception is the work Laura Klaming and Anton Vedder have done regarding the possibility of technological enhancement in the case of eyewitness testimony and memory. See Laura Klaming and Anton Vedder, "Brushing Up Our Memories: Can We Use Neurotechnologies to Improve Eyewitness Testimony?," *Law, Innovation, and Technology* 1, no. 2 (2009): 203–221; and "Human Enhancement and the Common Good: Using Neurotechnologies to Improve Eyewitness Memory," *AJOB Neuroscience* 1, no. 3 (2010): 22–33.

SOPHIE STAMMERS is a Research Fellow in the Department of Philosophy, University of Birmingham, ERI Building, Edgbaston, Birmingham B15 2TT, United Kingdom; e-mail <s.stammers@bham.ac.uk>. Her research interests include implicit cognition and confabulation.

traditionally or commonly used and that the theoretical implications of their use have not been extensively examined.[4] In the following, I discuss two interventions of particular relevance to cognitive bias mitigation that may emerge with considerable technological advances. These are technologies of a futuristic neuroscience, but we should not delay debate about their desirability for bias mitigation until they are fully developed: that conversation is worth having now.

1.1 Interfacing Cognitive Processes with Computers

Many forms of interfacing with computers with a view to enhancing cognition are reasonably mundane and commonplace: in fact, I am engaging in one such case of interfacing right now, by arranging and rearranging threads of sentences using a word processing program with the aim of developing the structure of my argument, a process that would be much more arduous on paper. All sorts of computer and smart phone applications provide potential interfacing opportunities that can facilitate and enhance cognition. However, those who think *proper* interfacing requires an especially tight connection — preferably between neural networks and silicon circuits — will be pleased to hear that researchers are making some relevant inroads on that front. One avenue of research demonstrates that implants can read neural signals so that a computer program may visualize the activity of the neural network in question.[5] Researchers are still some way from interpreting the *content* of representations across the network, but some envisage being able to do this one day.[6] It is conceivable, then, that the connection between mind and computer could become seamless, with the possibility that cognition is enhanced significantly through this direct integration with software.

1.2 Direct Neural Manipulation

Various forms of direct neural manipulation feature prominently in science fiction. For instance, in the film *The Matrix,* one of the characters directly "downloads" the know-how required to pilot a helicopter.[7] Meanwhile, *Eternal Sunshine of the Spotless Mind* features technology that enables people who have suffered heartbreak in a relationship to locate and delete all of their memories of the lover who broke their heart.[8] Such technology, at least as regards the breadth

4. This concern is shared by others; see, for example, Bostrom and Sandberg, "Cognitive Enhancement," 312–313.

5. See Miguel Nicolelis et al., in "Chronic, Multisite, Multielectrode Recordings in Macaque Monkeys," *Proceedings of the Academy of Sciences of the United States of America* 100, no. 19 (2003): 11041–11046; see also Jose Carmena et al., "Learning to Control a Brain–Machine Interface for Reaching and Grasping by Primates," *PLoSBiology* 1, no. 2 (2003): 193–208.

6. Miguel Nicolelis and Sidarta Ribeiro contemplate this in "Multielectrode Recordings: The Next Steps," *Current Opinion in Neurobiology* 12, no. 5 (2002): 602–606.

7. *The Matrix*, film, directed by the Wachowski Brothers (Burbank, CA: Warner Bros, 1999).

8. *Eternal Sunshine of the Spotless Mind*, film, directed by Michel Gondri (New York: Focus Features, 2004).

and accuracy possible in interventions of this sort, is still a long way off. But that doesn't mean that the implications of such interventions are not worthy of discussion. Moreover, recent developments in neuroscience suggest the beginnings of what Jiangyuan Hu calls "selective erasure," which may be of particular interest for our purposes in this article[9] By blocking a specific protein, Hu and colleagues were able to reverse particular kinds of associations formed in memory. Further research led by Dheeraj Roy demonstrates that lost memories can be reconstructed by manipulating engram cells (suspected sites of memory storage).[10] It is worth noting that the subjects in Hu et al.'s study were snails, while those in Roy et al.'s study were mice. However, both groups of researchers are confident that their respective methods could, in the future, be employed by humans, particularly in therapeutic applications (for treating anxiety or post-traumatic stress disorder, and Alzheimer's Disease respectively). We will discuss the implications of these findings in more detail in section 3.

Let us now turn our attention to some systematically epistemically faulty cognitions that these interventions may serve to improve.

2. Cognitive Bias and Knowledge Acquisition

Cartesian views in which our perceptual and cognitive systems typically represent reality as it is have generally fallen out of favor.[11] For some, this undermines the possibility of any accurate representation,[12] although others have countered this view, arguing that cognitive heuristics that result in some biased judgments develop for the reason that they deliver accurate representations *overall*.[13]

There is no commonly agreed upon definition of cognitive bias. This is perhaps because there is much discussion over what it means for a cognition to be "biased." To be biased is to deviate from some prescribed (set of) norms(s), but which set of norms is at issue is hotly debated.[14] For this reason, I prefer the definition offered by Martie Haselton and colleagues: it considers biases *distortions* of objective reality, without pronouncing on which norms they violate. According to Haselton et al., cognitive biases are "cases in which human cognition reliably produces

9. Jiangyuan Hu et al., "Selective Erasure of Distinct Forms of Long-Term Synaptic Plasticity Underlying Different Forms of Memory in the Same Postsynaptic Neuron," *Current Biology* 27, no. 13 (2017): 1888–1899.

10. Dheeraj Roy et al., "Memory Retrieval by Activating Engram Cells in Mouse Models of Early Alzheimer's Disease," *Nature* 531, no. 7595 (2016): 508–512.

11. See Andy Clark, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* (New York: Oxford University Press, 2016).

12. See, for example, Hans-Georg Gadamer in *Truth and Method*, trans. Joel Weisheimer and Donald G. Marshall (New York: Continuum, 1989).

13. See, for example, Gerd Gigerenzer and Wolfgang Gaissmaier in "Heuristic Decision Making," *Annual Review of Psychology* 62, no. 1 (2011): 451–482.

14. For discussion, see Andrea Polonioli, "Adaptive Rationality, Biases, and the Heterogeneity Hypothesis," *Review of Philosophy and Psychology* 7, no. 4 (2016): 787–803.

representations that are systematically distorted compared to some aspect of objective reality."[15] Cognitive biases also affect how humans seek, interpret, and form judgments about incoming information, resulting in further distortions.[16]

Let us turn to two commonly recognized types of cognitive bias: confirmation bias (2.1) and social bias (2.2). I show why each is of particular relevance to educational settings and justify the need to discuss whether and how we should eliminate them through technological interventions should the possibility arise.

2.1 Confirmation Bias

People are more likely to search for, and to accept, information that conforms with their existing beliefs and hypotheses than that which contradicts them. This tendency alone might not always deliver distorted cognitions or distortion of the relevant features of the decision environment: If one's existing beliefs correspond with reality, accepting information that coheres with these may well result in the adoption of beliefs that do correspond with reality. However, most of us will, in many circumstances, harbor some beliefs that do not correspond with reality or, at least, that do not capture all of the relevant factors in a decision environment. In these cases, confirmation bias is apt to shape our thinking.[17]

Confirmation bias threatens learning in disciplines concerned with how things are in the world and that center on creating models that correspond with reality (for example, the sciences and humanities). Take, for example, belief in the reality of anthropogenic climate change. The only considerations that ought to determine whether humans are causing increasing global temperatures and a changing climate are climate data. However, it turns out that existing political beliefs play a role in whether people accept the existence of anthropogenic climate change. People who are politically conservative are significantly more likely to believe that anthropogenic climate change is not happening.[18] The effect is present in those with increased levels of science comprehension,[19] and is even seen among (non–climate) scientists,[20] meaning that not just learners, but, potentially,

15. Martie G. Haselton, Daniel Nettle, and Paul W. Andrews, "The Evolution of Cognitive Bias," in *The Handbook for Evolutionary Psychology*, ed. David M. Buss (Hoboken, NJ: John Wiley, 2005), 968.

16. For discussion of different cognitive biases, see Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).

17. Raymond Nickerson, "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* 2, no. 2 (1998): 175–220.

18. Aaron M. McCright and Riley E. Dunlap, "The Politicization of Climate Change and Polarization in the American Public's Views of Global Warming, 2001–2010," *Sociological Quarterly* 52, no. 2 (2011): 155–194.

19. Dan M. Kahan et al., "The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks," *Nature Climate Change* 2, no. 10 (2012): 732–735.

20. J. Stuart Carlton et al., "The Climate Change Consensus Extends beyond Climate Scientists," *Environmental Research Letters* 10, no. 9 (2015): 94025.

educators (assuming at least some of the scientists in question also have teaching duties) are susceptible to confirmation biases in their practice.

Confirmation bias in the above instance has been explained in part by people's motivation to avoid inconsistency: people who are politically conservative tend to support free-market economics and are averse to market restrictions on commodities like fossil fuels. However, it is uncomfortable to admit that one champions something that has harmful results, and more comfortable to reject the existence of the harms in question. The risk that our values might determine, in part, which information we find persuasive and credible presents considerable risk to many aspects of education (including primary, secondary, further, and higher education): from determining the content included (and omitted) from curricula (for example, omitting meaningful study of colonialism from UK history curricula because it does not cohere with one's notion of British values); to the evaluation of student work and provision of feedback (heavy marking workloads entail considerable time pressure, which increases cognitive load — ideal conditions for the manifestation of biases such as confirmation bias). As such, if we are interested in providing rigorous, comprehensive, and applicable educational experiences, we ought to look into mitigating confirmation bias.

2.2 Social Bias

People who profess to having egalitarian commitments, and who do not intend to discriminate, have nonetheless been shown to make unfavorable judgments about a person and/or his or her accomplishments on the basis of that person's social group membership. Researchers explain these findings by proposing that we tend to associate certain social identity groups with certain stereotypical traits and, further, that these associations can manifest in cognition and lead us to misrepresent people by expecting them to comply with a stereotype that does not necessarily reflect reality.

Researchers have described these as *implicit* social biases, although there is much discussion regarding exactly what is implicit about them, as well as how they are best measured.[21] Social biases arise in a variety of teaching and learning environments.[22] They are of particular relevance to educational theorists and practitioners, from an epistemic perspective, because they cause us to discount the scholarly contributions of people who do not fit the social stereotype associated

21. For an overview of the controversy surrounding (some) measurement paradigms, see Jesse Singal, "Psychology's Favorite Tool for Measuring Racism Isn't Up to the Job," *The Cut* (*New York Magazine*), January 11, 2017, http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html); and for a series of responses reinstating the existence of social bias and the need to combat it, see John Schwenkler, "What Can We Learn from the Implicit Association Test? A Brains Blog Roundtable," The Brains Blog, January 17, 2017, http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx.

22. Guy Boysen and David Vogel, "Bias in the Classroom: Types, Frequencies, and Responses," *Teaching of Psychology* 36, no. 1 (2009): 12–17; and Cheryl Staats, "Understanding Implicit Bias: What Educators Should Know," *American Educator* 39, no. 4 (2015–2016): 29–43.

with their field of knowledge, which could lead to distorted representations of academic progress in that field.

According to one study, for example, people evaluate an error-ridden piece of writing less favorably when they believe it was written by an African American person compared to when they think the author is white.[23] In this study, researchers presented participants with writings that were identical except for the name of the author, sometimes using names commonly associated with African American people, and sometimes using names commonly associated with white people. They found that participants perceived manuscripts by (apparently) African American authors as less accurate than those by (apparently) white authors. Researchers maintain that seeing the author's name activates racial stereotypes that distort participants' perception of the manuscript's quality. Effects like these are important to educational theorists and practitioners in light of evidence that there is a larger attainment gap between ethnic minority and non–minority students in classes taught by educators who demonstrate a higher degree of implicit prejudice versus those taught by educators with lower levels of implicit prejudice.[24] Devaluing intellectual contribution on the basis of social group membership is also shown to occur in academic research. One study demonstrated that university faculty across 259 institutions were more likely to reply to emails purporting to be from students requesting mentoring on a future research project if the email appeared to come from a white man rather than from a woman and/or ethnic minority correspondent.[25] Again, the researchers conducting this study kept the email content fixed, changing only the name of the sender. In another case, a journal showed a significant increase in the publication of articles by female authors following the introduction of double-blind review, in which the social identity of the author is unknown to those involved in the editorial process.[26] Taken together, these studies seem to confirm that bias leads to a discounting of cognitive contributions at least partially on the basis of social group membership. Such implicit social bias undermines our pursuit of knowledge, particularly when these contributions, if given appropriate credence, have the potential to further human inquiry and extend our understanding of the world.

23. Arin N. Reeves, "Written in Black & White: Exploring Confirmation Bias in Racialized Perceptions of Writing Skills" (Chicago: Nextions Yellow Paper Series, 2014), http://nextions.com/wp-content/uploads/2017/05/written-in-black-and-white-yellow-paper-series.pdf.

24. Linda van den Bergh et al., "The Implicit Prejudiced Attitudes of Teachers: Relations to Teacher Expectations and the Ethnic Achievement Gap," *American Education Research Journal* 47, no. 2 (2010): 497–527.

25. Katherine Milkman et al., "What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway into Organizations," *Journal of Applied Psychology* 100, no. 6 (2015): 1678–1712.

26. Amber Budden et al., "Double-Blind Review Favors Increased Representation of Female Authors," *Trends in Ecology and Evolution* 23, no. 1 (2008): 4–6.

2.3 Good Candidates for Technological Intervention

There is much discussion over exactly how the cognitions involved in the biases discussed in the preceding sections should be characterized. Theorists sometimes identify them as *unconscious*, suggesting that we are unaware that the relevant judgments are distorted by cognitive biases, taking our attitudes to correspond well with reality. However, research suggests that in some circumstances people might become aware of the influence of a bias on their behavior.[27] Thus, it might be better to think of them as typically unnoticed, but potentially available upon reflection. Recognition of bias in one's attitudes does not guarantee that a person will successfully overwrite the biased cognition immediately or prevent its influence in reasoning, however. Such biases are typically difficult to control;[28] at the very least, they require reasonably effortful strategies to be brought under control.[29] Furthermore, as Lisa Bortolotti demonstrates, we have a tendency to believe we are reasoning well, even when confronted with inconsistencies in our cognition. Bortolotti draws upon on a range of studies demonstrating our tendencies toward preference reversal and decision procedure variation.[30] For instance, altering whether one and the same disease management program is described in terms of lives saved as opposed to lives lost alters peoples' endorsement of that program.[31] However, evidence suggests that people can be very resistant to changing their position when confronted about their preference variability and other inconsistencies in these sorts of decisions.[32]

That recognition and control of bias requires effort and persistence does not mean that that people should not be expected to make the relevant efforts to bring biased cognition under control (as Jules Holroyd and Daniel Kelly argue).[33] Still, it seems appropriate to inquire as to whether any other strategies might make the process easier and more effective, and thus to turn our attention to whether technological interventions could help out.

Before moving on to the next section, it is worth outlining the scope of my argument and acknowledging other issues that I do not have space to address

27. See studies from Margo Monteith et al., "Taking a Look Underground: Detecting, Interpreting, and Reacting to Implicit Racial Biases," *Social Cognition* 19, no. 4 (2001): 395–417; and Adam Hahn, Charles M. Judd, Holen J. Hirsch, and Irene V. Blair, "Awareness of Implicit Attitudes," *Journal of Experimental Psychology: General* 143, no. 3 (2014): 1369–1392; and discussion by Jules Holroyd, "Implicit Bias, Awareness, and Imperfect Cognitions," *Consciousness and Cognition* 33 (May 2015): 511–523.

28. Neil Levy, *Consciousness and Moral Responsibility* (Oxford: Oxford University Press, 2014).

29. Jules Holroyd and Daniel Kelly, "Implicit Bias, Character, and Control," in *From Personality to Virtue*, ed. Alberto Masala and Jonathan Webber (Oxford: Oxford University Press, 2016).

30. Lisa Bortolotti, *Delusions and Other Irrational Beliefs* (Oxford: Oxford University Press, 2009), chap. 2.

31. Amos Tversky and Daniel Kahneman, "The Framing of Decisions and the Psychology of Choice," *Science* 211, no. 4481 (1981): 453–458.

32. Bortolotti, *Delusions and Other Irrational Beliefs*, 87.

33. Holroyd and Kelly, "Implicit Bias, Character, and Control."

in this article, but that have been explored more fully elsewhere. As previously highlighted, I am interested in why we might be motivated to investigate technological interventions on cognitive biases from an epistemic perspective — that is, in order to improve the process of knowledge acquisition and sharing in educational settings. This raises the question of the sense in which people are thereby obligated to utilize technological interventions when available in educational settings and invites a discussion of consent. For the purposes of this article, I will assume that consent from the individual undergoing the intervention would be a requirement for proceeding.[34] It is consistent with exploring technological interventions on cognitive biases from an epistemic perspective that other concerns (for example, those from a moral perspective) might independently motivate this investigation, but I will not consider those here.[35]

Others have cautioned that because novel technologies intended for enhancing any human capacity will come at some financial cost, their use will introduce unfair advantages for those able to afford these interventions, thereby disadvantaging those who cannot. I follow others in seeing these issues as symptomatic of a general problem of unfair resource distribution that is perpetuated under dominant economic systems rather than a particular issue for novel technological interventions.[36] While there is only space here to recommend technological interventions from an epistemic perspective, someone taking forward these recommendations would want to consider these practical issues more fully.

3. Epistemic Considerations for Technological Interventions on Cognitive Biases

In this section, I demonstrate that there are two ways in which we might use technology to intervene on cognitive biases, and then assess the desirability of these interventions for facilitating knowledge acquisition. The first kind of intervention acts on the processes that produce distorted representations. The second acts on individual representations. We will consider them in that order.

3.1 Halting Biased Processes

Recall the discussion from section 1 in which technologies of a future neuroscience can interpret and interact with the representation of information across a neural network. This might occur via sophisticated neural-interfacing implants (as

34. One might argue that the same considerations that obligate people in positions of power (for example, employers) to take implicit bias training might also obligate their use of novel technological interventions, and that this also applies to educators. But a fuller discussion would be necessary before proceeding.

35. For a discussion of using technological enhancement to extend our moral capacities, see Julian Savulescu and Ingmar Persson, "The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity," *Journal of Applied Philosophy* 25, no. 3 (2008): 162–177.

36. For discussion, see Bostrom and Sandberg, "Cognitive Enhancement."

in 1.1) that are able to scan the network's activity at a fine enough grain that software interfacing with the network can decode the representations and processes running on it. Let's also suppose that the interfacing device is able to manipulate representations and to halt, redirect, or initiate processes (as in 1.2).[37] Now, we can program our software to detect any process on the network that has resulted in a distorted representation (such as when the network utilizes a heuristic, or jumps to a conclusion that isn't properly supported by other representations in the network) and, via the interface, redirect the network to run processes that will result in a nondistorted representation instead.

Call the intervention described above the "Heuristic Terminator." By intervening, halting, and redirecting the processes which generate cognitive biases, the Heuristic Terminator ensures that the neural network will not produce the distorted representations which, as we saw above, can thwart our acquisition of knowledge. Use of the Heuristic Terminator would thereby significantly enhance knowledge acquisition in a person who previously demonstrated a typical degree of cognitive bias, enabling them to avoid the distorted cognitions discussed in section 2.

In principle, this all sounds good. However, once we start thinking more about how the Heuristic Terminator's software would work, together with the limitations of the existing "wet-ware," we run into some trouble. As we saw in 2.1, heuristic reasoning does not always result in distorted cognition. Sometimes it leads to further beliefs that *do* correspond with reality. Thus, it is not the case that terminating any heuristic process will kill off all and only the distorted cognitions. It may also result in the extinction of *accurate* representations.

It will be helpful at this point to consider why we tend to rely on heuristics and other cognitive shortcuts in the first place. Human information processing capacities are limited, but the information in our environment that is possibly relevant to cognition is extensive and far outweighs our processing capacity. Using heuristics enables us to more easily identify information that might be relevant and, moreover, enables us to terminate inquiry without having to consider every possibly relevant piece of information. It is likely that a vast quantity of human information processing relies on heuristics.[38]

It is unclear whether the Heuristic Terminator could differentiate a heuristic that typically delivers accurate cognitions from one that does not. If its instruction to the neural network is to terminate *all* heuristic processes and to demand that all representations are arrived at in a way that guarantees their reliability, then we might end up with a system that simply stops forming any further representations because the threshold for establishing an accurate representation of the world is

37. We do not yet have a story to tell about the details here, but, as suggested in section 1, we should not have to wait for an innovation to be available before debating whether — and how — it is worth utilizing.

38. See Thomas Gilovich et al., eds., *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge: Cambridge University Press, 2002), for a series of essays on this topic.

too high to be met with existing cognitive resources. In that scenario, one would regularly find oneself unable to form new beliefs. So, interfacing with the Heuristic Terminator would likely be a rather uncomfortable and disorientating experience. We have not arrived at a place where knowledge acquisition is facilitated — quite the opposite. The Heuristic Terminator has created the ultimate skeptic.

It might be suggested that if the problem is that human processing capacities are limited, then the solution is to use the Heuristic Terminator in conjunction with *another* technological intervention that significantly augments processing capacity. This could occur through further interfacing in which processing is offloaded from the low-capacity wetware and distributed across a much higher capacity artificial processor. Then, we would have a substantially upgraded processor to support the operations necessary for forming accurate representations without relying on heuristic processes.

Using this intervention, we would be able to form new beliefs while avoiding cognitive biases. However, one might now wonder exactly how much human is left in this extensive labyrinth of interfacing.[39] Suffice to say, this option constitutes a somewhat radical enhancement. It might be that in the future, humans will regularly use interfaces for a range of applications, and so this proposal will be more mundane for them than it seems to us now. Still there is another concern: the solution we have created here is rather inelegant. In ensuring that cognitive biases do not arise in the system, we have had to remove a significant part of what the system does *well*, only to bolt on a nonhuman module to meet the demand necessary to replicate those epistemically useful processes. Perhaps there is a less roundabout — and less radical — way to achieve the same effects.

3.2 Targeted Interventions

If heuristic processes in general bring epistemic benefits as well as costs, then perhaps a better arrangement than the Heuristic Terminator is an intervention that targets just those representations which, when recruited in processing, typically result in distorted cognitions. In fact, some cognitive scientists employ current technological interventions with the aim of achieving this targeted effect. A canonical example comes from a pioneer of research into implicit cognition, Mahzarin Banaji, who has made a screensaver that cycles through a thousand pictures of counterstereotypical images of people, with two main aims: (a) to combat stereotypical associations between concepts that are overemphasized by biased sources (for example, a media that overemphasizes the association between black people and criminality);[40] and (b) to give her access to representations of people and their

39. It might be thought that the discovery that heuristics are embedded in our cognition already undermines the essence of human thought, but it is not clear that this follows: it has been argued that such processes are, in fact, fundamental to interpreting our experiences and to building our understanding of the world (see, for example, Gadamer, *Truth and Method*).

40. A report on representations of black men in the media by the Opportunity Agenda found that negative associations tend to be exaggerated, while positive associations tend to be limited. The Opportunity

life experiences beyond the scope of her normal perspective. Her intention is that (a) enables her cognitive system to gradually uncouple stereotypical associations; while (b) allows her to instill new representations about people she would otherwise know little about, so that she is less likely to rely on inaccurate stereotypes of their life and experiences.[41] Speaking of these sorts of interventions, Banaji says, "I no longer believe that I can just let information into my mind as it comes. I believe I must choose and edit. … I actually am pleased that the way technology now allows me to craft what I want to watch and listen to."[42] It has been pointed out that these sorts of interventions require continued effort and, even then, may not rid someone completely of a distorted representation. Neil Levy maintains that it remains "controversial" as to whether interventions like those suggested by Banaji enable the overwriting of distortions in a manner that is "relatively rapid" or "arduous, slow and extremely uncertain."[43] So, one might be interested in whether future technologies may enhance this effortful and uncertain process.

As introduced in 1.4, Hu and colleagues' developments in "selective erasure," which aim to target particular associations in memory, might be the basis of one such intervention. As before, we can imagine that we are operating under a somewhat futuristic neuroscience, in which neural manipulation technology can search for specific associations between concepts, and can then selectively erase them, leaving accurate information regarding the concepts in question intact. Let's also imagine that our intervention is able to overwrite or implant new representations (as in Roy et al., also discussed in 1.4). Call this intervention the "Selective Manipulator." Let's now consider how it might work.

Recall, Reeves's findings that people judge one and the same piece of writing to be more error-ridden when they believe it to have been written by an African American person rather than by a white person.[44] By selectively erasing the association linking African Americans and academic underachievement, the Selective Manipulator could prevent distorted judgments of the accuracy of writing perceived to be by African Americans from occurring in this case. The Selective Manipulator could also be used to target associations between women and academic underachievement to prevent the kind of distorted judgments discussed in section 2.2.[45]

Agenda, "Social Science Literature Review: Media Representation and Impact on Lives of Black Men and Boys" (New York: The Opportunity Agenda, 2011), 13–14, http://racialequitytools.org/resourcefiles/Media-Impact-onLives-of-Black-Men-and-Boys-OppAgenda.pdf.

41. This technique was used in Nilanjana Dasgupta and Anthony Greenwald, "On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals," *Journal of Personality and Social Psychology* 81, no. 5 (2001): 800–814.

42. Interview of Mahzarin Banaji, "The Mind Is a Difference-Seeking Machine," *On Being with Krista Tippett* (podcast), June 9, 2016, https://onbeing.org/programs/mahzarin-banaji-the-mind-is-a-difference-seeking-machine-aug2018/.

43. Levy, *Consciousness and Moral Responsibility*, 99.

44. Reeves, "Written in Black & White."

45. Those revealed in Milkman et al., "What Happens Before?"; and Budden et al., "Double-Blind Review Favors Increased Representation of Female Authors."

The Selective Manipulator will be useful when it enables the deletion of a biased association that might otherwise become active in cognition to produce biased judgments. For instance, consider two teachers who harbor an association between the concepts *male* and *rationality* who are grading the quality of students' arguments. For the teacher whose association remains intact, that association is apt to distort their perception of argument quality (leading them to mark one and the same argument as of a higher quality when they believe it was written by a boy as compared to a girl), while this risk is removed for the teacher whose association has been selectively deleted. One might think that simply anonymizing student work produces the same result, but knowing who produced which piece of work while marking is pedagogically valuable, for it enables teachers to tailor the tone of their feedback (some students do better with frank, straightforward feedback, but for others this tone is not constructive).

We have so far considered associations between social identity and particular aptitudes. But what about cases in which teachers favor contributions that reflect the dominant culture and undervalue contributions that support nondominant cultures? For instance, in a politics class, a student raised with Western values, with an emphasis on individualism, might praise the individualistic aspects of a political system, while a student raised in a culture that places more emphasis on collectivism might criticize those aspects of a political system, and a Western teacher might unfairly undervalue the second student's contribution. While there is much empirical work on how associations regarding dominant social identity categories bias cognition, there is less on how dominant ideologies produce bias in cognition, and so it would be premature to make any claim regarding how the Selective Manipulator would work in these cases. If it turns out that dominant ideologies are, like social stereotypes, upheld through a series of discrete evaluations (in which, for example, facets of individualism are positively valanced while those of collectivism are negatively valanced), then it is possible that these may also be targeted by the Selective Manipulator. It could be that there is less public agreement on whether these cases count as bias because they may well be viewed through the lens of the dominant ideology; therefore, communication around such cases would need to be handled with care. But public pushback against current de-biasing efforts is common, and managing this is another practical issue to be considered before using the Selective Manipulator.[46]

Not everyone may support the use of the Selective Manipulator. For instance, those who espouse the "mirror view" of the above attitudes might object that this use of the Selective Manipulator will not have the intended epistemically beneficial outcomes. According to the mirror view, social biases are not really biases at all, but reflect real-life propensities. Nilanjana Dasgupta, for instance, suggests that implicit attitudes are "mirror-like reflections of local environments and communities within which individuals are immersed," and that "[t]hrough repetition, these observations get passively recorded in the mind and become

46. I thank an anonymous reviewer for pushing me to consider these sorts of cases.

the basis of implicit attitudes and beliefs."[47] For instance, African Americans *do* underachieve in some educational settings as compared with white people, and it is this fact that causes an association with underachievement, but social, economic, and political factors (such as reduced access to education and financial resources) explain this underachievement.[48] Nonetheless, for proponents of the mirror view, reality is reflected in a cognition that links African Americans with underachievement.[49] Following the mirror view, Tamar Szabó Gendler has argued that implicit social attitudes aren't really *biases* at all because they reflect real-world propensities and, further, that in rejecting them, one loses important accurate representations of the world.[50] If this is right, then from an epistemic point of view, these cognitions ought to be preserved after all.

However, a number of other philosophers have argued that the situation is more nuanced than the mirror view would have it. Alex Madva, for instance, maintains that the mirror view is "a radically oversimplified and misleading gloss on the psychology of prejudice," pointing to evidence that our social representations of the world are partly reinforced and maintained by the way we want to see the world.[51] For instance, one study shows that men were more critical of findings indicating a bias against hiring women in science, while women were more critical of findings indicating an absence of such a bias.[52] In short, confirmation bias, driven by the desire to, for instance, downplay the structural benefits that have given one advantages over members of another group, feeds and maintains our implicit social biases.

Further, one might hold something like a mirror view of implicit associations, but argue that use of the Selective Manipulator is nevertheless epistemically recommended. Katherine Puddifoot maintains that even if our implicit social attitudes do reflect real-world propensities (for example, associating science with men simply because there are more prominent male scientists than female

47. Nilanjana Dasgupta, "Implicit Attitudes and Beliefs Adapt to Situations: A Decade of Research on the Malleability of Implicit Prejudice, Stereotypes, and the Self-Concept," *Advances in Experimental Social Psychology* 47 (2013): 240–241.

48. Indeed, Tyrone Howard demonstrates both this differential achievement and that multiple factors (including the prejudiced attitudes and differential treatment by educators) account for these findings, in "Who Really Cares? The Disenfranchisement of African American Males in PreK–12 Schools: A Critical Race Theory Perspective," *Teachers College Record* 110, no. 5 (2008): 954–985.

49. See also Sally Haslanger, "Social Structure, Narrative, and Explanation," *Canadian Journal of Philosophy* 45, no. 1 (2015): 1–15.

50. Tamar Szabó Gendler, "On the Epistemic Costs of Implicit Bias," *Philosophical Studies* 156, no. 1 (2011): 33–63.

51. Alex Madva, "A Plea for Anti-Anti-Individualism: How Oversimple Psychology Misleads Social Policy," *Ergo* 3, no. 27 (2016): 719. These motivations might be unconscious, or at least relatively unreflective, and available only through careful self-observation.

52. Ian Handley et al., "Quality of Evidence Revealing Subtle Gender Biases in Science Is in the Eye of the Beholder," *National Academy of Sciences* 112, no. 43 (2015): 13201–13206, cited in Madva, "A Plea for Anti-Anti-Individualism," 719.

scientists), their tendency to feature in so much other processing that results in further distorted cognition outweighs the epistemic benefit of reflecting real-world propensities.[53] For instance, an association between black people and underachievement may reflect reality, but it is apt to be activated automatically by stimuli evoking black people and to generate distorted judgments in instances where this association should not have any normative force. Even if it is true that black people tend to underperform in some academic pursuits compared with people from other racial backgrounds, that is not a reason for seeing more errors in one and the same piece of writing when it is associated with a black author compared to a white author — and yet, that is what people believe they see, constituting a real risk for perpetuating stereotypes in educational settings.[54] For this reason, Puddifoot argues, the epistemic benefits of maintaining the association are outweighed by these downstream epistemic costs, and so we would do better epistemically if we rejected the association. Accordingly, the epistemic considerations that motivate the mirror view (reflecting reality) may still be compatible with the use of the Selective Manipulator.

While I am convinced by Madva's and Puddifoot's arguments that the epistemic characteristics of implicit biases are not exhausted by pointing to the ways in which they reflect society, I am also sympathetic to a concern raised by many who espouse the mirror view: that the relevant social stereotypes are connected to deep and pervasive structural issues that we should not lose sight of in discussions of cognitive bias.[55] *This* concern may well count against the use of the Selective Manipulator as a method to extinguish implicit social biases in favor of advancing learning — or, at least, it requires that the Manipulator be used in conjunction with other resources to mitigate this potential. If the Selective Manipulator allows users to effectively delete their biases without acknowledging their content, their source, or their part in perpetuating structural injustices, then it takes away an important opportunity to engage learners and educators with the aim of facilitating their recognition of the structures that constrain the trajectories of knowledge acquisition.

Consider again Reeves's finding that a piece of writing is evaluated more harshly when participants believe the author is African American than when they think the author is white.[56] In these sorts of judgments, the *manifestation* of the association is inappropriate, but participants *harbor* this association in

53. Katherine Puddifoot, "Dissolving the Epistemic/Ethical Dilemma over Implicit Bias," *Philosophical Explorations* 20, sup. 1 (2017): 73–93.

54. Reeves, "Written in Black & White."

55. Haslanger, "Social Structure, Narrative, and Explanation." There is disagreement about the nature of the connection: those who espouse the mirror view think injustice in the structure of society is primarily responsible for the formation of stereotypes in cognition, while some respondents to the mirror view (including Madva) think that there is feedback, with the cognitive stereotypes reinforcing unjust structures, and those structures reinforcing the stereotypes.

56. Reeves, "Written in Black & White."

part because African American students generally do underachieve in comparison with white peers — yet, this attainment gap exists due to historical structural injustices that deprive African American communities of resources necessary to develop academic success.[57] The association itself does not contain information about the historical structural injustice that explains the achievement differential. Nevertheless, confronting the fact that one harbors negative associations about African Americans presents an opportunity to deepen one's understanding of these structural issues. This could prove particularly important for teachers. In this case, it contextualizes and justifies the provision of support to African American students. Extinguishing the relevant association through application of the Selective Manipulator removes this pedagogically significant opportunity.

Even Sally Haslanger, who cautions against expending too many philosophical resources on the discussion of the cognitive aspects of bias, points out that "drawing attention to implicit bias can be strategically useful as a starting point for discussion of social injustice because there is empirical evidence to support the claim that we are all biased."[58] But with the Selective Manipulator, the opportunity for discussion does not necessarily arise, and so we miss out on discussing important structural issues.

This potential outcome does not require that we forgo any use of the Selective Manipulator — it still might be the most effective method for preventing the distorted cognitions that thwart educational goals, as discussed in 2.2. But we should act to preserve the opportunity to turn attention to unjust structures that freely accompanies traditional implicit bias interventions. In order to preserve these pedagogically important opportunities, I propose that the Selective Manipulator be designed such that it both requires users to confront the content of their biases, as well as how they have figured in cognition, and simultaneously provides them with information about the wider social and historical context, inviting them to engage with it and consider how it relates to their biases.

One might think that this suggestion is motivated by egalitarian concerns rather than by epistemic concerns; however, greater attention to the structures that reinforce implicit social biases not only works toward egalitarian goals, but also facilitates achieving the *epistemic* goals discussed in 2.2, in turn promoting better teaching and learning. If we wish to see progress in various fields of knowledge acquisition and dissemination, we should not erroneously discount the scholarly contributions of people who do not fit the social stereotype associated with a particular field of knowledge. Structural barriers at the heart of mirror theorists' concerns also create barriers to development in the field of knowledge. Amia Srinivasan has argued forcefully for this point regarding progress

57. Howard, "Who Really Cares?"

58. Haslanger, "Social Structure, Narrative, and Explanation," 12.

in philosophy, for instance. She maintains that all philosophy is value-laden.[59] Even in logic and proof theory, we appeal to values such as simplicity and elegance, for instance. Moreover, which values are designated as desirable within a discipline is in part a matter of cultural and historical standards. According to Srinivasan,

> once we recognise that the outputs of our philosophical theorising are radically shaped by how, where and with whom we are thrown into the world, then we will see the philosophical pressure to diversify philosophy. A homogenous discipline means a homogenous set of ideas, a homogenous set of intellectual products and projects. If our goal is to collectively explore logical space, collectively seek the truth, then a genealogically homogenous search party won't be particularly good at the job.[60]

In other words, disallowing people to participate in pursuing and extending knowledge acquisition on the basis that they do not fit the dominant social stereotype of that discipline is not a neutral action as regards progress in that discipline. Scholars who have homogenous social and cultural experiences may advance their discipline in fewer directions than a more heterogeneous workforce.[61]

One may think that this is only true in select subjects — science is not value-laden, for instance, or so the reply goes. But that isn't so clearly true. Even in scientific disciplines, scholars make decisions about which phenomena are worthy of scientific investigation; which discoveries are worthy of acknowledgment and attention in the standardized canon; and which examples from the canon are taught. For example, in medicine, a historically male discipline, the contemporary research agenda still proceeds along a gendered trajectory.[62] The sciences are disciplines that purport to offer *comprehensive models* of reality, not reality as modeled from the perspective of some subset of society, and so diverse experiences are integral to setting the course for how, and into what, scientific inquiry proceeds. As such, from an epistemic perspective (as well as from a moral one), we ought to be aware of and interested in dismantling the structures that perpetuate marginalization and prevent diversification in places of research, teaching, and learning.

59. Amia Srinivasan, "Does Feminist Philosophy Rest on a Mistake?" (keynote address presented at the King's College London Minorities and Philosophy Conference, July 4, 2015); transcript available at http://users.ox.ac.uk/~corp1468/Research_files/Does%20Feminist%20Philosophy_KCL%20talk.pdf

60. Ibid.

61. It is worth clarifying that Srinivasan's view is not relativism about accurate understanding of the world, but rather that our epistemological project to achieve accurate understanding is more effectively served when it draws on the participation of people with different experiential starting points.

62. For instance, there is a significant lack of research into endometriosis, as well as into uterine fibroids — both gynecological conditions — compared to other diseases with similarly debilitating symptoms. See Geoffrey Adamson et al., "Creating Solutions in Endometriosis: Global Collaboration through the World Endometriosis Research Foundation," *Journal of Endometriosis* 2, no. 1 (2010): 3–6; and L. Amanti et al., "Uterine Leiomyoma and Its Association with Menstrual Pattern and History of Depo-Medroxyprogesterone Acetate Injections," *International Journal of General Medicine* 4 (2011): 535–538.

Conclusion

In the foregoing, I argued that because cognitive biases bring about distorted cognitions, they thwart goals of knowledge acquisition and dissemination, and therefore educators have a strong interest in their mitigation. Future technological interventions have the potential to enhance bias mitigation. I argued that an intervention that halts the processes that lead to distorted cognitions, the Heuristic Terminator, will not serve to facilitate knowledge acquisition because heuristic processes often result in accurate cognitions, and without them we would find it very difficult to make sense of decision environments. For this reason, a more targeted approach, the Selective Manipulator, was favored; it works by erasing particular associations that lead to distorted cognitions.

I defended the use of the Selective Manipulator against a challenge from the mirror view, but argued that we should take seriously the idea that structural barriers to participation are worthy of our attention and showed how they can hinder the progression of knowledge expansion and dissemination. Since moral and epistemic aims converge on engendering an understanding of structural injustice, use of the Selective Manipulator presents an opportunity to host a public conversation about the origins of the relevant biases in the unjust structures that they also act to uphold. Accordingly, if the Selective Manipulator is to be used for the purpose of facilitating those epistemic goals discussed in 2.2, we ought to ensure that we preserve the opportunity for understanding structural injustices.

The discussion here has focused mostly on social biases, but there is a broader lesson. Technological interventions such as direct neural manipulation take us out of a more organic learning environment where inquiry can take any number of tangents and allow one to make new discoveries that in turn lead to epistemic progression. When future technological interventions promise cognitive shortcuts, they may serve to facilitate specific epistemic goals. But we should also consider what we might have learned if we had gone the long way around and, if necessary, factor this into our use of the technological intervention.

THE AUTHOR ACKNOWLEDGES THE SUPPORT of a European Research Council Consolidator Grant, Pragmatic and Epistemic Role of Factually Erroneous Cognitions and Thoughts (PERFECT), Grant Agreement 616358.