CANADIAN JOURNAL OF PHILOSOPHY, 2016
VOL. 46, NO. 3, 369–386
http://dx.doi.org/10.1080/00455091.2016.1176982

Routledge
Taylor & Francis Group

The reversal test, status quo bias, and opposition to human cognitive enhancement

Steve Clarke[a,b]

[a]Centre for Applied Philosophy and Public Ethics, Charles Sturt University, Canberra, Australia; [b]Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK

ABSTRACT

Bostrom and Ord's reversal test has been appealed to by many philosophers to substantiate the charge that preferences for status quo options are motivated by status quo bias. I argue that their characterization of the reversal test needs to be modified, and that their description of the burden of proof it imposes needs to be clarified. I then argue that there is a way to meet that burden of proof which Bostrom and Ord fail to recognize. I also argue that the range of circumstances in which the reversal test can be usefully applied is narrower than they recognize.

ARTICLE HISTORY Received 4 November 2015; Accepted 7 April 2016

KEYWORDS Argument from risk; backfiring; burden of proof; human cognitive enhancement; reversal test; status quo bias

1. Introduction

An increasingly popular way for philosophers to argue against opponents who prefer status quo options, in this or that domain of inquiry, is to try to explain away their preferences as resulting from 'status quo bias.' Dorsey (2010) suggests that welfare-relevant preferences are often shaped by status quo bias. De Brigard (2010) suggests that preferences for real experience over virtual experience, as expressed in ordinary reactions to Nozick's experience machine thought experiment, are often motivated by status quo bias. Wilkinson (2009) suggests that much of the opposition to opt-out approaches in medicine is driven by status quo bias. Kahane and Savulescu (2015) suggest that opposition to human enhancement is motivated by status quo bias, and there are many more examples of explanatory appeals to status quo bias that can be found in the recent philosophical literature.

CONTACT Steve Clarke stclarke@csu.edu.au

The mere fact that someone prefers the status quo to change does not show that they are in the grip of status quo bias. Sometimes, in some circumstances, the status quo option is the best available choice and would, therefore, be the choice of an unbiased, rational person. To substantiate the charge that a particular preference results from status quo bias, we need to offer supporting evidence over and above a mere demonstration that status quo bias might possibly have shaped that preference. In a much cited article, Bostrom and Ord (2006) present a test, which, they argue, can be used to provide corroborating evidence for the presence of status quo bias.[1] This is the reversal test. The reversal test has been appealed to by many, including Wilkinson (2009, 236), Sandberg (2011, 83), Rippon (2012, 350–354), and Kahane and Savulescu (2015, 138).[2]Bostrom and Ord describe the reversal test as follows:

> *Reversal Test*: When a proposal to change a certain parameter is thought to have bad overall consequences, consider a change to the same parameter in the opposite direction. If this is also thought to have bad overall consequences, then the onus is on those who reach these conclusions to explain why our position cannot be improved through changes to this parameter. If they are unable to do so, then we have reason to suspect that they suffer from status quo bias. (2006, 664–665)

The reversal test can be used to substantiate the assertion that a preference for a status quo option, which can be represented as the value of a parameter, results from status quo bias. According to Bostrom and Ord,

> The rationale of the Reversal Test is simple: if a continuous parameter admits of a wide range of possible values, only a tiny subset of which can be local optima, then it is prima facie implausible that the actual value of that parameter should just happen to be at one of these rare local optima.

Therefore, they argue, '… the burden of proof shifts to those who maintain that some actual parameter is at such a local optimum' (2006, 665). The reversal test shifts the burden of proof by showing up the *prima facie* irrationality of favoring the status quo without providing good reason for doing so. Bostrom and Ord (2006) discuss four ways in which one might go about finding reasons that are sufficient to meet that burden of proof: appealing to evolutionary adaptation, appealing to transition costs, and appealing to arguments that they refer to as 'the argument from person-affecting ethics' and 'the argument from risk' (2006, 665–70).

Despite the widespread application of the reversal test, Bostrom and Ord's (2006) discussion of it has been subjected to very little criticism, and what little it has been subjected to is mostly indirect criticism.[3] Nordmann (2007) is critical of the way in which Bostrom and Ord (2006) appeal to status quo bias to explain opposition to human cognitive enhancement.[4] He argues that Bostrom and Ord (2006) encourage their readers to assume that they must accept either that humans have a nature that just happens to be at optimal parameter values, or that humans should be cognitively enhanced (Nordmann 2007, 38–39). But this is a false dichotomy. Nordmann (2007) understands the reversal test

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

to play a part in Bostrom and Ord's (2006) setup for this false dichotomy, but it only plays a part and it is not a necessary one. Nordmann would make the same criticism of their appeal to status quo bias to explain opposition to human cognitive enhancement even if it did not involve the reversal test. His major line of criticism of Bostrom and Ord (2006) is not a criticism of the reversal test. He does not argue that all uses of the reversal test encourage acceptance of the false dichotomy he warns us about, and he does not dispute that the reversal test can be an effective test of status quo bias.[5]

Nebel (2015) also discusses the reversal test. The main focus of his paper is status quo bias. Whereas most people, including, it appears, Bostrom and Ord (2006), simply assume that those who are in the grip of status quo bias are less than fully rational, Nebel (2015) argues that status quo bias can be consistent with full rationality. In particular, he argues that on some theories of the rationality of preferences, especially (but not only) on subjective theories, it is not always irrational to be biased toward the status quo. Nebel is a proponent of the reversal test. But while Bostrom and Ord (2006) present the reversal test as a test for status quo bias, Nebel holds that the reversal test should be understood as a test for *irrational* status quo bias (Nebel 2015, 453–455).[6] As we will see, the reversal test works by identifying instances of status quo bias which are, *prima facie*, irrational, so the reversal test tests for both status quo bias and irrational status quo bias.

None of the authors who have critically discussed, or applied the reversal test appear to stop to consider whether and when it is an effective test of status quo bias, and whether it might not be refined. My concern in this paper is with these unaddressed issues. Whether or not the reversal test is best characterized as a test for status quo bias (which, following Bostrom and Ord (2006) will be my working assumption), or irrational status quo bias, will not matter for my purposes. I'll take issue with Bostrom and Ord's (2006) wording of the reversal test and with their characterization of the burden of proof it imposes. I'll show that there is a way to try to meet that burden of proof which Bostrom and Ord do not consider. I'll demonstrate that they do not succeed in showing that opponents of human cognitive enhancement – their key example – fail to meet the burden of proof imposed by the reversal test; and more generally, that they overestimate the range of cases in which the reversal test can be usefully applied. I will not, however, dispute that the reversal test can sometimes be used to show up the irrationality of preferring status quo options, and can therefore be an effective test of both status quo bias and irrational status quo bias.

In the next section of the paper, I'll briefly explain why psychologists and cognitive scientists, such as Kahneman (2011), are convinced that ordinary thinking is prone to status quo bias. In the following section, I'll explain how Bostrom and Ord (2006) think that the burden of proof imposed by the reversal test can be met. I'll also introduce the key example that Bostrom and Ord (2006) use in their discussion of the reversal test: opposition to human cognitive enhancement

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

(2006, 656). In the fourth, fifth, and sixth sections I discuss the problems that I discern in Bostrom and Ord's (2006) discussion of the reversal test. Section Seven contains concluding remarks. Throughout the discussion, I'll assume Bostrom and Ord's definition of status quo bias: '… an inappropriate (irrational) preference for an option because it preserves the status quo' (2006, 658).

2. Status quo bias

Psychologists and cognitive scientists who work on 'status quo bias' do not claim that there is a single factor that biases our thinking in favor of the status quo. Rather, they hold that a persistent bias in our thinking, toward the status quo, results from the combined influence of a number of distinct psychological effects. One of these is the pervasive 'endowment effect,' which leads people to ascribe additional value to items when they come to regard these as their possessions. To exchange what we possess for an alternative involves departing from the status quo, so by deterring people from making exchanges, the endowment effect causes status quo bias.[7]

Another contributor to status quo bias is loss aversion. Many of the opportunities we are presented with are 'mixed,' involving the possibility of gain and the risk of loss. It seems that when we decide to accept or reject such opportunities, we typically do not weigh potential losses and potential gains equally. We feel the pain of a loss more than we feel the satisfaction of a gain. So, we tend to avoid mixed opportunities unless these are markedly in our favor (Kahneman 2011, 284). Our tendency to forego opportunities that are only marginally in our favor leads us to avoid making many changes to our circumstances which we would otherwise make and thereby contributes to status quo bias.[8] There are also various other psychological effects that contribute to status quo bias including omission bias (Ritov and Baron 1992) and regret avoidance (Anderson 2003).[9]

3. The burden of proof imposed by the reversal test and opposition to human cognitive enhancement

In Section Five, I will dispute Bostrom and Ord's claim that in order to meet the burden of proof imposed by the reversal test good reason needs to be provided to suppose that we are currently at a local optimum with respect to the value of a parameter (2006, 665). However, they are surely right to think that good reasons for preferring the current value of a parameter to changes in either direction should be required of the proponent of the status quo. Status quo bias provides a *prima facie* explanation of a preference for the current value of a parameter, so the reasons for preferring the status quo to changes in either direction should be the basis for an explanation that is at least as good an explanation as a competing explanation of a preference for the current value of a parameter that involves appeal to status quo bias. Otherwise, the most plausible conclusion

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

to draw is that status quo bias is the best available explanation of a preference for the current value of a parameter.

Bostrom and Ord (2006) work through the example of opposition to human cognitive enhancement, in some detail, to illustrate the application of the reversal test. They ask us to suppose that a medically safe, affordable means of enhancing human cognition has been developed. They then consider whether or not this new technology should be used. They set aside deontological constraints and ask, 'do we have reason to believe that the long-term consequences of human cognitive enhancement would be, on balance, good?' (2006, 656). Their answer is yes, but, as they note, many people's answer is no (Bostrom and Ord 2006, 662–663). Many believe that there are unacceptable risks involved in allowing cognitive enhancements to be used, even if these are medically safe and affordable. It has been suggested that in a world in which inheritable genetic alterations were widely available, humanity might split into two sub-species, a post-human sub-species and an unenhanced sub-species; and that members of the post-human sub-species might enslave or exterminate unenhanced humans (Annas, Andrews, and Isasi 2002, 154).[10] It has also been suggested that enhanced humans might lack a property that Sandel (2004) refers to as 'openness to the unbidden.' And it has been suggested that in a world in which people could be genetically enhanced, the ability of parents to determine the genes of their children might lead to various undesirable consequences, as parents are liable to make shortsighted choices on behalf of those children (Fukuyama 2002, 93–100).[11]

The long-term consequences of human cognitive enhancement might be deleterious, all things considered, but it is also possible that they will be beneficial for us, overall. Nick Bostrom imagines that overall beneficial consequences have been obtained from cognitive enhancement in his 'Letter from Utopia.' In this letter 'your possible future self' writes from a future in which post-humans enjoy a highly fulfilling life, every second of which '… is so good that it would blow our minds had their amperage not been previously increased' (Bostrom 2010, 8). Bostrom and others in the pro-enhancement camp do not deny that there is some chance that things will not turn out as they hope, and do not deny that the widespread use of cognitive enhancement techniques could lead to one or more of the undesirable outcomes that opponents of enhancement warn us about. Similarly, most of those in the anti-enhancement camp do not deny that there is some chance that the negative outcomes that they worry about will not transpire, or will not be as worrisome as they anticipate. The disagreement is over competing estimates of the likely overall balance of costs and benefits arising from human cognitive enhancement.

If we could perform an accurate cost-benefit analysis, determining the various harms that human cognitive enhancement is liable to cause, and the likelihood that these will occur, and weigh this against a determination of the various benefits that human cognitive enhancement is liable to provide, along with accurate

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

estimates of the likelihood that these will occur, then perhaps we could settle the debate between opponents and proponents of human cognitive enhancement. But we cannot do this. It is very difficult to determine the harmfulness of some of the potential hazards of human cognitive enhancement. How harmful is a loss of openness to the unbidden? It's very hard to say. Also, we cannot currently acquire accurate information about the likelihood that cognitive enhancement will lead to the various benefits and harms that have been mooted occurring. People's current convictions about the likelihood of these occurring are the result of intuitive judgments; and these are grounded on incomplete information and subject to cognitive biases (Gilovich, Griffin, and Kahneman 2002; Kahneman 2011). This is not a point that Bostrom and Ord (2006) dispute. They describe assessment of the overall consequences of human cognitive enhancement, as well as other assessments of consequences that impact on 'human lives and social systems,' as judgments that '… rely also – crucially and unavoidably – on subjective, intuitive judgment' (2006, 657). Despite acknowledging the crucial role that subjective, intuitive judgment plays in assessment of the overall consequences of human cognitive enhancement, Bostrom and Ord assure us that application of the reversal test shows that objections to human cognitive enhancement are affected by status quo bias and that 'When the bias is removed, the objections are revealed as extremely implausible' (2006, 658).

Opponents of human cognitive enhancement are hardly ever in favor of cognitive dis-enhancement. So, it seems that they are committed to defending the view that average human cognitive abilities just happen to be at a local optimum, which seems prima facie unlikely, given all the different values at which average human cognitive abilities might be found. So, the onus is upon the defender of the status quo, with respect to human cognitive abilities to demonstrate that we are at a local optimum. Bostrom and Ord attempt to apply all four of the possible ways of meeting the burden of proof imposed by the reversal test, which they identify, to the case of opposition to human cognitive enhancement (2006, 665–672). They argue that none can be successfully applied to this case. I'll go on to dispute the conclusion of this argument in Section Five, but first I'll raise an issue about their wording of the test and consider a way of meeting the burden of proof it imposes that Bostrom and Ord (2006) do not consider.

4. Backfiring and the reversal test revised

Bostrom and Ord's question: 'do we have reason to believe that the long-term consequences of human cognitive enhancement would be, on balance, good?' (2006, 656) is ambiguous. They might be asking us to imagine a future in which many humans have been cognitively enhanced to a significant degree and then asking us to try to judge whether or not life in this future would be preferable, all things considered, to life now. Or they might be asking us to imagine all the

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

possible consequences of allowing significantly many humans to attempt to become cognitively enhanced, and then asking whether, when we weigh the likelihood of any or all of these consequences eventuating, we think it is better to allow, or to try to prevent, human cognitive enhancement? On the second interpretation of their question, a form of risk will need to be considered that it is not necessary to consider on the first interpretation. This is the risk that our attempts to enhance ourselves could fail and cause unintended deleterious outcomes – which is to say that they could backfire – and lead to a future that is neither one in which cognitive enhancements are widely used, nor one that is superior to the status quo.

Bostrom and Ord do not consider the possibility of backfiring, so it seems reasonable to understand them as construing their question narrowly. But it is possible that our attempts to create a cognitively enhanced future for ourselves will backfire and this possibility concerns some opponents of human enhancement. We can try to use the reversal test to investigate whether or not intuitions about the significance of this possibility are driven by status quo bias, but it will require revisions to the wording of the test for us to be able to do so. These revisions should be made anyway, because any attempt to shift the value of a parameter in a particular direction could backfire, and defenders of the status quo will be able to appeal to the possibility of backfiring to defend the status quo in many different contexts. A reversal test that fails to make room for this possibility is significantly less useful than one that does.

Bostrom and Ord's wording of the reversal test includes the phrase: 'When a proposal to change a certain parameter is thought to have bad overall consequences, consider a change to the same parameter in the opposite direction' (2006, 664). Their wording is, in effect, an invitation to consider the consequences of a proposed change to a parameter that has succeeded, along with the consequences of a change that has taken place in the opposite direction. What we should be trying to consider are all the possible consequences of attempting to change a parameter in one direction, and all the possible consequences of attempting to change it in the other direction, including the possibilities of attempts to change the parameter in either direction failing. Here is a revised version of the reversal test that allows for the possibility that attempts to change the value of a parameter may fail (while taking account of earlier revisions), and so is broad enough to allow for consideration of the possibility of backfiring. New wording has been underlined:

> *Reversal Test* (Revision One): When an attempt to change a certain parameter is thought to have bad overall consequences, consider an attempt to change the same parameter in the opposite direction. If this is also thought to have bad overall consequences, then the onus is on those who reach these conclusions to explain why our position cannot be improved through attempts to change this parameter. If they are unable to do so, then we have reason to suspect that they suffer from status quo bias.

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

The propensity of human schemes aimed at improving society to backfire is a persistent theme in conservative political thought (Buchanan 2011, 146). Given that many of the opponents of human enhancement are politically conservative, it should not be surprising that some are concerned about potential backfiring. Both Fukuyama (2002, 3–10) and Kass (2003, 11) worry that attempts to enhance humans may inadvertently lead to the creation of a society that resembles the dystopic future society depicted in Aldous Huxley's *Brave New World* ([1932] 2007). This looks to be a concern about an attempt to create an utopian society backfiring and leading to the creation of a society in which a future world government uses scientific techniques to cognitively dis-enhance large sections of the population.

Brave New World is a society controlled by a world government in which selective breeding, and interventions in fetal development, lead to the creation of distinct human castes – Alphas, Betas, Gammas, Deltas, and Epsilons – bred to play specific roles in a world economy. The developing fetuses of Gammas, Deltas, and Epsilons are cognitively dis-enhanced by being deprived of oxygen and doused with alcohol to ensure that they have artificially low levels of intelligence ([1932] 2007, 4–11). The sperm and ova that is used to create Alphas and Betas is selected to ensure that their intelligence is at a suitable level for their designated economic roles, but there is no attempt to manipulate the genetic material of these castes to produce higher-than-natural levels of intelligence mentioned in the book. *Brave New World* is far from the scientifically advanced world that Nick Bostrom imagines in his *Letter from Utopia* (2010). Scientific advances may have led to *Brave New World*, but it is a world with a government that prohibits serious scientific research, on the grounds that scientific breakthroughs could threaten social stability (Huxley [1932] 2007, 198).

Neither Fukuyama (2002) nor Kass (2003) try to explain how the cognitive enhancement of humans might backfire, leading to a Brave New World-style scenario in which systematic cognitive dis-enhancement becomes commonplace; and Huxley himself does not spell out how he supposes that our world might become the dystopia of his imagination.[12] Perhaps Fukuyama and Kass fear that after the means to alter human cognitive capacities is developed, it may fall into the hands of a malevolent sub-group of humanity who use it not to enhance anyone, but as part of a scheme to create an oligarchic world government, dis-enhancing the cognitive capacities of the people who might object in the process?

Could an argument against human cognitive enhancement, based on fears that attempt to cognitively enhance humans will inadvertently lead to a Brave New World-style dystopic future, be adapted to become an argument that meets the (revised) reversal test? To apply the (revised) reversal test we need to consider the consequences of attempts to either increase or decrease the cognitive powers of significantly many humans. If it is considered reasonable to fear that attempts to enhance cognitive capacity will lead to a dystopic future, in which humans fall under the control of a world government that manipulates individual cognitive capacities, then it also seems reasonable to fear that attempts to

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

dis-enhance cognitive capacity could lead to a similar dystopic future involving a world government manipulating individual cognitive capacities. This might involve successful cognitive dis-enhancement, as in *Brave New World*, or it might involve attempts to dis-enhance backfiring, but still leading to power being concentrated in the hands of a malevolent oligarchy. So, it looks like at least one of the arguments mounted against cognitive enhancement, based on the fear of backfiring, can be adapted to become an argument for the cognitive status quo, which looks like it might meet the (revised) reversal test. More generally, it seems that any attempt to shift the value of a parameter away from the status quo is susceptible to the possibility of backfiring, and appeals to this possibility have the potential to meet the burden of proof imposed by the (revised) reversal test.

5. The argument from risk

Bostrom and Ord introduce the reversal test and immediately specify what needs to be done to meet the burden of proof that it imposes: provide 'some good reason' to think that we are at a local optimum with respect to a parameter (2006, 665). They then go on to sketch four ways in which one might meet the burden of proof imposed by the reversal test. When they come to discuss meeting that burden of proof by employing the 'argument from risk,' it becomes apparent that Bostrom and Ord (2006) are willing to consider a way of meeting that burden that does not involve providing good reason to suppose that we are at a local optimum with respect to a parameter. They write:

> Even if it is agreed that we are probably not at a local optimum with respect to some parameter under consideration, one could still mount an argument from the risk against varying the parameter. If it is suspected that the potential gains from varying the parameter are quite low and the potential losses very high, it may be prudent to leave things as they are. (2006, 669)

However, they go on to assert, mere uncertainty about consequences is not good grounds for sticking with the status quo, 'Only if the expectation value of the hypothetical negative results is larger than the expectation value of the hypothetical positive results does the uncertainty favor the preservation of the status quo.' (2006, 668–669).

What Bostrom and Ord (2006) appear to be suggesting is that if it can be shown that the potential harms of increasing the value of a parameter outweigh the potential benefits, and if it can be shown that the potential harms of decreasing the value of a parameter outweigh the potential benefits, then we have sufficient reason to retain the status quo; and we do not need to also demonstrate that we are at a local optimum with respect to a parameter. So, in effect, they are proposing a less demanding burden of proof than the requirement that we demonstrate that we are at a local optimum with respect to the value of a parameter. Bostrom and Ord (2006) appear to intend that this less demanding burden of proof only be applied when a version of the 'argument

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

from risk' is employed. But the less demanding burden of proof seems appropriate to apply whenever we are employing the (either revised or unrevised) reversal test. If we can demonstrate that the potential harms of attempting to increase the value of a parameter outweigh the potential benefits, and the potential harms of attempting to decrease the value of a parameter outweigh the potential benefits, then we have explained why our position cannot be improved through changes to a parameter, which is what the reversal test asked of us (Bostrom and Ord 2006, 664–665).

Bostrom and Ord's (2006) less demanding burden of proof is highly relevant to the discussion of their key example, opposition to human cognitive enhancement. Many opponents of human cognitive enhancement think that the potential harms of attempts to cognitively enhance humans outweigh potential benefits; and they take it as given that the potential harms of attempting to cognitively dis-enhance humans outweigh potential benefits. If they are able to demonstrate that they are right about both of these claims, then it looks like they will be able to meet the burden of proof imposed by the (either revised or unrevised) reversal test, without also having to demonstrate that human cognitive capacity fortuitously happens to be at a (locally) optimal level.

I'll look at Bostrom and Ord's (2006) discussion of the risks involved in human cognitive enhancement in some detail in the next section, but before I do so I want to raise another issue about their wording of the reversal test that is particularly relevant to a discussion of the argument from risk. Their wording of the reversal test places the onus on those who favor the status quo with respect to a parameter 'to explain why our position cannot be improved through changes to this parameter' (2006, 664). But appeal to the argument from risk to meet the burden of proof imposed by the reversal test does not involve showing that our position *cannot* be improved by attempting to make changes to a parameter, it involves showing that our position is more likely to be made worse than improved through attempting to make changes to the parameter (in either direction). Therefore, the reversal test should be revised as follows. Again, changes from the original wording are indicated via underlining:

> *Reversal Test* (Revision Two): When an attempt to change a certain parameter is thought to have bad overall consequences, consider an attempt to change the same parameter in the opposite direction. If this is also thought to have bad overall consequences, then the onus is on those who reach these conclusions to explain why our position is more likely to be made worse than improved through attempts to change this parameter. If they are unable to do so, then we have reason to suspect that they suffer from status quo bias.

6. The risks involved in cognitively enhancing humans

As we saw earlier, Sandel (2004), Annas, Andrews, and Isasi (2002), and Fukuyama (2002) all argue, in different ways, that we should refrain from cognitively enhancing humans because of the risks of deleterious consequences. They all

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

offer arguments against human cognitive enhancement, rather than arguments for the status quo, with regard to human cognitive ability, but their respective arguments could easily be transformed into arguments for the status quo, by being conjoined with arguments against human cognitive dis-enhancement. Human cognitive dis-enhancement is usually regarded as a very unattractive option, so such arguments should not be hard to provide.

As we have already noted, Bostrom and Ord (2006) suggest that what is needed to determine whether or not a particular instance of the 'argument from risk' meets the burden of proof imposed by the reversal test is an assessment of the overall consequences of departing from the status quo with respect to the value of a parameter. There will be many contexts in which we will be unable to provide fully objective assessments of these consequences and will have to settle for somewhat subjective ones. Assessments of the potential costs and benefits involved in human cognitive enhancement are such cases. According to Bostrom and Ord, as we saw earlier, our assessments of all risks to 'human lives and social systems' rely on subjective intuitions about risks, which cannot be objectively weighed (2006, 657).

Having explained that the way to assess the argument from risk is to weigh all of the potential costs and benefits of human cognitive enhancement and dis-enhancement, one might expect Bostrom and Ord to attempt an assessment of the overall costs and benefits of human cognitive enhancement and dis-enhancement, but they do not do this. Instead, they present us with an argument for the conclusion that we may have systematically underestimated the expected benefits of human cognitive enhancement, as well as a listing of some of its most significant potential benefits (2006, 668–670). They are not clear about their motives, but it seems plausible to think that they intend to shift the balance of other people's assessments of the expected costs and benefits of human enhancement by presenting their argument and their listing.

Bostrom and Ord's argument for the conclusion that the benefits of human cognitive enhancement may have been systematically underestimated involves a thought experiment aimed at showing how hard it is for us to conceive of many of the potential benefits of cognitive enhancement (2006, 69). They ask us to consider the difficulties faced by a tribe of Australopithecus debating whether or not to enhance their intelligence to modern human levels. It would be extremely difficult for the Australopithecus to conceive of many of the benefits of human-level intelligence, such as having the abilities to create and appreciate art, literature, music, and poetry. Just so, they argue, we find it hard to foresee the various benefits of significantly enhanced cognition (Bostrom and Ord 2006, 669). The significant potential benefits of human cognitive enhancement they list are enabled by increased cognitive capacity. With increased cognitive capacity, we are more likely to be able to find cures to diseases, answer scientific questions, invent solutions for poverty, and solve environmental problems than we would be if our cognitive powers remain at their current level (Bostrom and Ord 2006, 669).

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

Having presented the above considerations, Bostrom and Ord acknowledge that some of their opponents will not accept that risks should be managed by weighing the potential costs and benefits of change (2006, 669). Some urge a risk-averse approach under which (consistent with loss aversion) potential costs weigh more heavily than benefits.[13] Bostrom and Ord (2006) think that even if we adopt such a risk-averse approach the case for human cognitive enhancement is still strong. This is because, they argue, increased human cognitive capacity can enable us to reduce various serious threats to humanity more effectively than we would be able to otherwise (2006, 670). All things being equal, we are more likely to find cures to new pandemic diseases, more likely to be able to prevent asteroids from colliding with the Earth, more likely to be able to prevent invasion by hostile aliens from other solar systems, and so on, if we are cognitively enhanced than if we are not.

Bostrom and Ord's discussion of cognitive biases that are liable to lead us to underestimate the likely benefits of human cognitive enhancement is relevant to the question of whether or not opponents of human cognitive enhancement can meet the burden of proof imposed by the (revised) reversal test, however, cognitive biases cut both ways. Bostrom and Ord's debating Australopithecus would have a hard time conceiving of the benefits of art, literature, music, and poetry. But they would also have a hard time conceiving of many of the disbenefits of modern life. For example, they would be unlikely to conceive of the various diseases that were not a threat to humans until we began to live in concentrated urban communities, and commenced farming, such as tuberculosis, smallpox, and influenza.[14] The debating Australopithecus would also have a hard time anticipating the many health problems caused by modern diets and would probably find some of the frustrating features of modern life, such as the difficulties involved in having to complete many years of formal education, having to dealing with modern bureaucracies, and having to function in a modern workplace, simply unimaginable. It's not clear that if our debating Australopithecus were able to comprehend all of the potential benefits and harms of having the intelligence of modern humans that they would choose to have their intelligence increased.[15] *Mutatis mutandis*, if modern humans were able to comprehend all of the potential benefits and harms involved in becoming cognitively enhanced post-humans.[16]

The other two lines of argument that Bostrom and Ord (2006) develop concern benefits to humanity that we are more likely to enjoy if we are cognitively enhanced than if we remain unenhanced, and harms to humanity that we are more likely to be able to avoid. Bostrom and Ord's (2006) discussion of these lines of argument is shot through with a sunny-eyed optimism about the propensity of cognitively enhanced individuals to devote their energies to providing benefits to, and reducing the risk of harms to, the rest of humanity. If we do not share this optimism then the conclusions that Bostrom and Ord (2006) draw look very shaky. Unless they are provided with the right incentives, selfish cognitively

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

enhanced people will not be motivated to cure diseases that do not endanger them, share scientific knowledge, alleviate the poverty of others, and protect parts of the environment that they do not interact with. Furthermore, selfish cognitively enhanced people may create and spread diseases that harm or kill others, prevent scientific knowledge from being shared, cause others to live in poverty, and cause damage to the environment, if it is in their interest to do these things. They can be relied on to use their enhanced intellectual powers to prevent pandemics, asteroid strikes, and alien invasions, when these place them in danger, but if they can protect themselves from such threats without also protecting others, then they are liable to do so.[17]

The result of the thought experiment involving the debating Australopithecus is inconclusive. The other two lines of argument for human cognitive enhancement can only be accepted as good arguments if we accept Bostrom and Ord's optimistic implicit assumption about the propensity of the cognitively enhanced to use their cognitive powers to aid the rest of humanity. As we have already noted, many of the opponents of human enhancement, such as Fukuyama (2002) and Kass (2003), are political conservatives and are likely to reject this optimistic assumption. Traditionally, political conservatives have taken a jaundiced view of human motivation, stressing the propensity of selfish motives and emotions to undermine pro-social tendencies (Buchanan 2011, 147).[18] So, many opponents of human cognitive enhancement are unlikely to be swayed by Bostrom and Ord's (2006) counterarguments to their arguments from risk against human cognitive enhancement. These opponents of human cognitive enhancement are unlikely to be able to persuade Bostrom and Ord, or other proponents of human enhancement, to shift their position either. The competing assessments that proponents and opponents of human cognitive enhancement make, about the likely balance of costs and benefits involved with attempts to cognitively enhance humans, are very different and because these assessments rely significantly on subjective intuition there is no clear way to adjudicate between them.

We lack accurate data to ground an assessment of the potential benefits of human cognitive enhancement. We also lack accurate data to ground an assessment of the potential costs involved with successful attempts to cognitively enhance humans, as well as the potential costs associated with attempts to cognitively enhance humans that backfire. Until such time as we have accurate data, to ground our assessments of all of the potential costs and benefits involved in attempts to cognitively enhance humans, proponents and opponents of human cognitive enhancement are going to have to continue to rely on their subjective intuition about costs and benefits, and disputes between the two sides are not going to be rationally resolved.

Do contemporary opponents of human cognitive enhancement do enough to meet the burden of proof imposed by the (revised) reversal test right now? Those who have pro-enhancement intuitions will be likely to conclude that

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

opponents of human cognitive enhancement do not do enough to meet the burden of proof imposed by the (revised) reversal test. However, those who have anti-enhancement intuitions are likely to think that at least some opponents of human cognitive enhancement succeed in meeting the burden of proof imposed by the (revised) reversal test. They will be convinced that the risks of harms resulting from (both successful and backfiring) attempts to cognitively enhance humans outweigh potential benefits[19]; and also convinced that the risks of harms resulting from (both successful and backfiring) attempts to cognitively dis-enhance humans outweigh potential benefits. Because disagreement about whether the burden of proof imposed by the (either revised or unrevised) reversal test has been met or not, comes down to competing subjective intuitions, it will not be amenable to rational resolution. Contra Bostrom and Ord (2006), application of the (either revised or unrevised) reversal test does not provide a clear result that tells against opponents of human cognitive enhancement and it does not reveal that objections to human cognitive enhancement are 'extremely implausible.' It produces different results depending on whether opponents or proponents of human cognitive enhancement run the test.

A general lesson we can take away from Bostrom and Ord's application of the reversal test to opposition to human cognitive enhancement is this: it is pointless to apply the (either revised or unrevised) reversal test to circumstances in which there is significant disagreement about the risks involved in varying the value of a parameter away from the status quo, and no prospects of rationally resolving that disagreement, as is the case when disagreement turns on competing subjective intuitions about those risks. In such circumstances, there cannot be an objective, impartial determination as to whether the burden of proof imposed by the test has been met or not.

7. Concluding remark

I have argued for four conclusions in this paper. First, that the wording of the reversal test should be revised as per the wording set out at the end of Section Five. Second, that the burden of proof needed to meet the (either revised or unrevised) reversal test is less demanding than that described by Bostrom and Ord (2006, 665). My clarification of the burden of proof needed to meet the reversal test was provided in Section Five. Third, that there is an additional way to meet the burden of proof imposed by the test, which Bostrom and Ord (2006) fail to recognize. This is by appealing to backfiring, as was discussed in Section Four. Fourth, the (revised) reversal test should not be applied to circumstances, such as opposition to human cognitive enhancement, where there is significant disagreement about the risks involved in varying the value of a parameter away from the status quo, and where that disagreement is not amenable to rational resolution. The argument for this conclusion was developed in Section Six.

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

I have been critical of Bostrom and Ord (2006). But despite the various shortcomings I have discerned in their (2006) paper, we should not lose sight of what Bostrom and Ord have achieved. The reversal test needs to be refined, and it needs to be applied more judiciously than they recognize, but nevertheless, it can be an effective test of status quo bias. Given the propensity of contemporary philosophers to try to explain away preferences for status quo options by appealing to status quo bias, it should be recognized that Bostrom and Ord (2006) have made an important contribution to philosophy by developing the original version of the much-used reversal test.

Notes

1. Bostrom and Ord (2006) has been cited over 144 times, as of 3 April 2016 (Googlescholar).
2. Bostrom and Ord (2006) also introduce a second test for status quo bias, the 'double reversal test.' This rather complicated test does not appear to have been applied by any philosophers, at least not in published articles and books. Criticisms of it have been made by Weidemann (2009, 125–7) and Nordmann (2007, 39, n. 22).
3. There is some critical discussion of reversal tests in a recent paper by Sparrow (2015) as well as in a set of published commentaries on that paper. However, Sparrow does not apply Bostrom and Ord's (2006) reversal test. Instead he applies his own test, which is loosely modeled on Bostrom and Ord's test. The differences between Sparrow's test and Bostrom and Ord's reversal test are discussed in two of the commentaries: Campbell and Wahlert (2015) and Powell (2015).
4. Nordmann also accuses Bostrom and Ord (2006) of failing to comprehend that they need to treat approaches to ethics that are neither consequentialist nor deontological seriously. See (Nordmann 2007, 39, n. 22).
5. Nordmann compares Bostrom and Ord's use of the reversal test, to reverse the burden of proof against the opponent of human cognitive enhancement, to contexts in which creationists illegitimately claim that creationism is on an evidential par with the theory of evolution, because biologists cannot offer absolute proofs of the truth of evolutionary theory (2007, 39). This comparison misses its mark. Creationists seek to evade the burden of proof that mainstream science imposes on those who promulgate theories that are inconsistent with accepted science. However, creationists do not seek to impose a burden of proof on evolutionary theorists. They merely seek equal treatment for creationism and evolutionary theory. So they seek to shift, rather than reverse, the burden of proof. Also, by treating evolution and creationism as just two unproven theories, the creationist evades consideration of the overwhelming evidential basis for acceptance of evolutionary theory. But there does not appear to be any analogously strong evidential basis for the rightness of opposition to human cognitive enhancement that Bostrom and Ord (2006) attempt to evade.
6. It is not clear that Nebel appreciates that he is redescribing Bostrom and Ord's reversal test when he describes it as a test for *irrational* status quo bias. See Nebel (2015, 453–455).
7. The endowment effect is nicely illustrated by a simple experiment due to Jack Knetsch. See Kahneman (2011, 296–297).

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

8. For discussion of a simple study, which vividly illustrates the influence of loss aversion on ordinary thinking, see Kahneman (2011, 283–284).
9. For a recent survey of effects contributing to status quo bias, see Eidelman and Crandall (2012).
10. Humans already enslave and exterminate other humans. Presumably, Annas, Andrews, and Isasi's (2002) underlying concern is that post-humans would be more effective enslavers and exterminators of humans than are humans.
11. References for some more suggested risks are collected by Bostrom and Ord (2006, 663, n. 14).
12. However, there is some discussion of this topic in Huxley ([1958] 2007).
13. There are many risk-averse approaches to the management of risk and many involve applying variants of the 'precautionary principle.' The precautionary principle is not often advocated explicitly in discussion of the right and wrongs of human enhancement, but there are commonalities between the forms of risk aversion advocated by many of the opponents of human enhancement and those advocated by proponents of the precautionary principle (Briggle 2014).
14. On the origins of urban diseases, see Diamond (2012, 295).
15. Modern lifespans are, on average, significantly longer than the lifespans of Australopithecus (Sacher 1975). If Bostrom and Ord's Australopithecus were to raise their intelligence to the level of modern humans then they would, presumably be able to figure out how to extend their lifespans, at least to some extent. This seems like a clear point in favor of enhancing their intelligence. It is hard to know how happy Australopithecus were. But available evidence suggests that members of pre-industrial cultural groups who lead uncomplicated lives are at least as happy as those living in modern industrialized societies (Biswas-Diener, Vittersø, and Diener 2005). A cognitively enhanced Australopithecus would, all things being equal, live a longer life than an unenhanced Australopithecus, but there seems to be no good reason to think that it would be a happier life.
16. Bostrom and Ord (2006) might concede that the various cognitive biases that infect ordinary reasoning would make it hard for the Australopithecus to appreciate the full extent of the dis-benefits of modern levels of intelligence, but go on to argue that cognitive biases might lead them to fail to weigh benefits and dis-benefits even-handedly; and so fail to appreciate the overall benefits of modern levels of intelligence. Given the large number of cognitive biases that infect human cognition, in different and conflicting ways, and which can combine to have unexpected effects, it seems that such a line of argument would be speculative at best.
17. A potential response to this line of objection would be to invoke the possibility of moral enhancement, as advocated by Persson and Savulescu (2012). If cognitively enhanced people were also morally enhanced, then their selfish and parochial tendencies would be replaced by pro-social tendencies. I don't think this is a credible response. I am persuaded by Powell and Buchanan (forthcoming) that an appreciation of the evolutionary history of human morality should lead us to conclude that proposals to use biological and technological interventions to morally enhance people are very unlikely to succeed.
18. For an extended discussion of this theme in conservative thought, see Kekes (1998, 68–90).
19. Some of the concerns raised by opponents of human cognitive enhancement are about the potential deleterious side effects of successful human cognitive enhancement. Others are about the potential deleterious effects of attempts to cognitively enhance humans backfiring. As far as I know, no opponent of human cognitive enhancement has explicitly distinguished between these two different forms of potential harm.

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

Acknowledgment

Thanks to Matthew Kopec, Neil Levy, Toby Ord, Russell Powell, and two anonymous referees for helpful comments on earlier drafts of this paper as well as audiences at the University of Oxford, Charles Sturt University, the University of Helsinki, and the Australasian Association of Philosophy annual conference in 2015.

Funding

This work was supported by the Australian Research Council under Discovery Grant [DP130103658].

Notes on contributor

Steve Clarke is an associate professor in the Centre for Applied Philosophy and Public Ethics, Charles Sturt University and a senior research associate of the Uehiro Centre for Practical Ethics at the University of Oxford. He is the author of *The Justification of Religious Violence*. Malden MA: Wiley-Blackwell, 2014. He is also a co-editor of Clarke, S., Powell, R. and Savulescu. J. eds. 2013. *Religion, Intolerance and Conflict: A Scientific and Conceptual Investigation*. Oxford: Oxford University Press; and Clarke, S., Savulescu, J., Coady, C. A. J., Giubilini, A. and Sanyal, S. eds. 2016. *The Ethics of Human Enhancement: Understanding the Debate*. Oxford: Oxford University Press.

References

Anderson, C. J. 2003. "The Psychology of Doing Nothing: The Effect of Imagining Behavioural Scripts on Personal Intentions." *Journal of Personality and Social Psychology* 45: 293–305.

Annas, G. J., L. B. Andrews, and R. M. Isasi. 2002. "Protecting the Endangered Human: Toward an International Treaty Prohibiting Cloning and Other Inheritable Alterations." *American Journal of Law and Medicine* 28: 151–178.

Biswas-Diener, R., J. Vittersø, and E. Diener. 2005. "Most People Are Pretty Happy, but There is Cultural Variation: The Inughuit, the Amish and the Massai." *Journal of Happiness Studies* 6: 205–226.

Bostrom, Nick. 2010. "Letter from Utopia." An earlier version was published in *Studies in Ethics, Law and Technology* 2 (1):1–7. Version 1.9. Accessed November 10, 2014. http://www.nickbostrom.com/utopia.pdf

Bostrom, Nick, and Toby Ord. 2006. "The Reversal Test: Eliminating Status Quo Bias in Applied Ethics." *Ethics* 116: 656–679.

Briggle, Adam. 2014. "Bioconservatism as Customised Science." In *The Customization of Science: The Impact of Political and Religious Worldview on Contemporary Science*, edited by S. Fuller, M. Stenmark and U. Zackariasson, 176–192. Basingstoke: Palgrave Macmillan.

Buchanan, Allen. 2011. *Beyond Humanity? The Ethics of Biomedical Enhancement*. Oxford: Oxford University Press.

Campbell, S. M., and L. Wahlert. 2015. "Is Disability Conservation Rooted in Status Quo Bias?" *American Journal of Bioethics* 15 (6): 20–22.

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press

De Brigard, Felipe, 2010. "If You like It, Does It Matter If It is Real?" *Philosophical Psychology* 23 (1): 43–57.
Diamond, Jared. 2012. *The World until Yesterday*. New York: Viking.
Dorsey, Dale. 2010. "Preference, Welfare and the Status-Quo Bias." *Australasian Journal of Philosophy* 88 (3): 535–554.
Eidelman, Scott, and Christian S. Crandall. 2012. "Bias in Favour of the Status Quo." *Social and Personality Psychology Compass* 6 (3): 270–281.
Fukuyama, Francis. 2002. *Our Posthuman Future*. New York, NY: Farrar, Straus and Giroux.
Gilovich, T., D. Griffin, and D. Kahneman. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.
Huxley, Aldous. [1932] 2007. *Brave New World*. London: Vintage.
Huxley, Aldous. [1958] 2007. *Brave New World Revisited*. London: Vintage.
Kahane, Guy, and Julian Savulescu. 2015. "Normal Human Variation: Refocussing the Enhancement Debate." *Bioethics* 29 (2): 133–143.
Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar Straus and Giroux.
Kass, Leon. 2003. "Ageless Bodies, Happy Souls: Biotechnology and the Pursuit of Perfection." *The New Atlantis*, Spring, 9–28.
Kekes, John. 1998. *A Case for Conservatism*. Ithaca: Cornell University Press.
Nebel, Jacob M. 2015. "Status Quo Bias, Rationality, and Conservatism about Value." *Ethics* 125 (2): 449–476.
Nordmann, Alfred. 2007. "If and Then: A Critique of Speculative NanoEthics." *NanoEthics* 1: 31–46.
Persson, Ingmar, and Julian Savulescu. 2012. *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press.
Powell, Russell. 2015. "The Disvalue of Genetic Diversity, or: How (Not) to Treat a Sandelian Ethos on Steroids." *American Journal of Bioethics* 15 (6): 29–32.
Powell, Russell and Buchanan, Allen. Forthcoming. "The Evolution of Moral Enhancement." In *The Ethics of Human Enhancement: Understanding the Debate*, edited by S. Clarke, J. Savulescu, C. A. J. Coady, A. Giubilini, and S. Sanyal, 239–260. Oxford: Oxford University Press.
Rippon, Simon. 2012. "How to Reverse the Organ Shortage." *Journal of Applied Philosophy* 29 (4): 344–358.
Ritov, I., and J. Baron. 1992. "Status Quo and Omission Biases." *Journal of Risk and Uncertainty* 5: 49–62.
Sacher, George A. 1975. "Maturation and Longevity in Relation to Cranial Capacity in Hominid Evolution." In *Primate Functional Morphology and Evolution*, edited by R. H. Tuttle, 417–442. The Hague: De Gruyter.
Sandberg, Anders. 2011. "Cognitive Enhancement: Uploading the Brain." In *Enhancing Human Capacities*, edited by J. Savulescu, R. ter Muelen and G. Kahane, 71–91. Oxford: Oxford University Press.
Sandel, Michael. 2004. "The Case against Perfection: What's Wrong with Designer Children, Bionic Athletes, and Genetic Engineering." *The Atlantic* 3 (293): 51–62.
Sparrow, R. 2015. "Imposing Genetic Diversity." *American Journal of Bioethics* 15 (6): 2–10.
Weidemann, Christian. 2009. "Towards a Heuristic for Nanoethics: The Onus of Proof in Applied Ethics. Uncovering Status Quo and Other Biases." In *Size Matters: Ethical, Legal and Social Aspects of Nanobiotechnology and Nano-Medicine*, edited by J. A. Ach and C. Weidemann, 117–131. Münster: LIT Verlag.
Wilkinson, Dominic. 2009. "Challenging the Status Quo." *Journal of Bioethical Enquiry* 6 (2): 235–237.

https://doi.org/10.1080/00455091.2016.1176982 Published online by Cambridge University Press