In Defence of ‘Toma’: Algorithmic Enhancement of a Sense of Justice¹

Biography

David Stevens is Associate Professor of Political Philosophy in the School of Politics and International Relations, University of Nottingham, UK. His work falls within the areas of analytical political philosophy, and rational choice theory. His main areas include religious views, and the philosophy of education. He is author of (with Kieron O’Hara) *The Devil’s Long Tail: Religious and other Radicals in the Internet Marketplace* (Oxford).

Abstract

Despite serious reservations over issues of transparency, accountability, bias, and the like, algorithms offer a potentially significant contribution to furthering human well-being via the influencing of beliefs, desires, and choices. Should governments be permitted to leverage socially beneficial attitudes, or enhance the well-being of their citizens via the use of algorithmic tools? In this chapter I argue that there are principled moral reasons that do not permit governments to shape the ends of individuals in this way, even when it would contribute a positive benefit to well-being. Such shaping would undermine the kinds of ethical independence that state legitimacy is based upon. However, I also argue that this does not apply to what Rawls calls a ‘sense of justice’ – the dispositions necessary to uphold just political and socioeconomic institutions. Where traditional methods of influence, such as education, prove lacking, then algorithmic enhancement towards those ends may be permissible. Mireille Hildebrandt’s fictitious piece of computational software – ‘Toma’ – serves as the point of departure for this argument, and provides many of the insights regarding the autonomic nature of such influence.

Key words

Algorithms; Rawls; political liberalism; a sense of justice; moral education; moral enhancement

Introduction

‘Toma’ is a piece of advanced computational software that employs algorithms to predict, structure, and alter the attitudes, behaviour, and choices of Diana, its human user (Hildebrandt 2015, 1-10).² Toma is dispersed across a number of platforms – handheld device, home, car, office computer system – and links with other Personal Data Assistants (PDAs), as well as third-party service providers, including government. Toma illustrates the potential that, in combination with Big Data, algorithms can significantly enhance human decision-making. Humans are “hackable”: by correcting for familiar flaws that beset human reasoning, algorithms can enable individuals to better fulfil their preferences, act in

¹ For discussions on the subject-matter of this essay I thank Matthew Clayton, Carlos Lopez-Benitez, Kieron O’Hara, Rebecca Orton, and Patrick Tomlin. For their most helpful comments on an earlier draft I thank Kieron O’Hara and, in particular, an anonymous reviewer.

² References to Hildebrandt 2015 are given as plain page numbers in text.

accordance with their values, or promote their well-being. Toma also balances the interests of Diana against the interests of third parties, potentially harmonising and optimising outcomes by shaping beliefs, desires, and choices.

Part of the Toma story illustrates the various concerns about algorithmic transparency, accountability, agency, and bias. In this chapter I set aside such concerns about the background conditions that might lead us to worry about the practical implications of employing algorithms, and concentrate instead on the positive transformative effects of algorithms for enhancing human well-being via the structuring of choices and behaviour. My focus, specifically, is on the normative question of what limits there are over the state's use of algorithms to steer our unconscious minds. In particular, I focus on the permissibility of the state using algorithms to influence the *moral* beliefs, attitudes, and behaviour of its citizens for socially beneficial ends. With governments increasingly interested in alternatives to traditional methods of gaining compliance with public policies (incentives and coercive threats), algorithms increasingly offer ways of bringing citizens to socially beneficial moral attitudes and behaviour (Wilkinson 2013, 341). Faced with the threat of a range of serious political issues that call for large-scale coordinated action – such as global warming – finding workable methods for gaining compliance is pressing (see Persson and Savulescu 2011; Rosa 2013). Increasingly sophisticated algorithms, that is, offer benefits in multiple directions: individuals benefit because their well-being is improved by optimising choices that will better achieve their own ends; society benefits by having its members adopt attitudes and behaviours that are conducive to peaceable social cooperation; and, taxpayers benefit by governments using less costly means to achieve the same ends.

Despite such positive benefits, there exist very real concerns about the manipulation that increasingly powerful algorithms will facilitate. Even when algorithms work perfectly, normative concerns remain that such manipulation is undermining of autonomy or dignity. The central question addressed here is whether it is morally permissible for the state to use this technology to interfere in society to promote individually and socially beneficial ends. In what follows, I argue for a two-fold conclusion: that there are principled reasons for the state to refrain from seeking to maximise the well-being of its citizens, but that it is permissible for the state to seek to develop in citizens a sense of justice that would underpin socially cooperative behaviour. If the state were to employ technologies such as algorithms to promote the well-being of its citizens – by shaping their goals, beliefs, and choices – then it would undermine the fundamental interest that citizens have in deciding for themselves what goals or ends to pursue. However, this kind of independence depends upon a scheme of social and political institutions that are only capable of being maintained if citizens possess the right kinds of attitudes regarding those institutions and the proper treatment of others; a sense of justice. Employing technology to shape these attitudes is, I argue, a morally permissible role of the state.

This chapter has the following format. I begin by exploring some of the transformative possibilities that algorithms promise. Second, I consider some familiar justifications that might block the state's use of algorithms to promote individual well-being. I reject these on the basis that they do not work in the case of algorithms. Third, I point to other grounds that might block state interference in the ends that citizens pursue on grounds that it undermines the freedom of

citizens to plan, revise, and pursue ends they choose for themselves, regardless of whether this promotes their well-being. Fourth, I develop this account in terms of political morality – about the proper relationship between citizen and state. Finally, I argue that although the state is not permitted to promote particular ends, it can employ technological enhancements to bring about attitudes and beliefs in citizens regarding a sense of social justice and cooperation.

The potential of algorithms for enhancement

Let us flesh out the Toma story. Diana employs Toma to help organise her life more efficiently and effectively. Toma is highly sensitive to the needs, moods, physical state, and preferences of Diana. Toma shapes and controls many features of Diana's world by filtering information, balancing competing claims on her time and attention, and adjusting her environment. Toma 'reads' Diana, anticipating her responses based on previous actions and a wealth of external data (pp.59-61). Toma's computer system is *autonomic*: unsupervised learning algorithms and neural networks that employ forms of artificial intelligence, advanced statistical methods, and machine learning (pp.23-26; Hildebrandt 2016). It is trained on, and works in tandem with, Big Data: vast quantities of data distributed across multiple datasets (pp.31-40; Pasquale 2015, 19-58; Hildebrandt 2013, 2-7). As Hildebrandt writes:

Our lifeworld is increasingly populated with things that are trained to foresee our behaviours and pre-empt our intent. These things are no longer stand-alone devices, they are progressively becoming interconnected via the cloud, which enables them to share their 'experience' of us to improve their functionality. We are in fact surrounded by adaptive systems that display a new kind of mindless agency (pp.viii-ix).

Toma goes beyond prediction. Toma exhibits a form of agency, actively altering Diana's beliefs, attitudes, and behaviour, enabling her to make better decisions in line with what it takes to be her wider interests. Toma operates at two levels: First, it influences the actual *choices* Diana makes by restricting the feasible option set, or making it more likely that a better option will be selected from the set. In this, algorithms ameliorate familiar flaws in human reasoning that cause individuals to depart from fully rational decision-making, such as myopia, framing, loss aversion, and overconfidence (see Persson and Savulescu 2011, 12-41; Savulescu and Maslen 2015, 80-81; Kahneman 2011, 278-374; Sunstein 2014, 8-13). In this, Toma performs tasks similar to 'nudging': structuring the background against which individuals can be led to make better (more optimal) choices, given their desires (see Sunstein 2014; Hausman and Welch 2010). Toma 'frames' choices, and 'primes' Diana.

Second, Toma operates directly on the very *beliefs* and *desires* Diana has. It filters information, and it induces or forestalls emotional responses by regulating various features of the environment or Diana's physiology that impact on her attitudes and decision-making. In this, it is akin to 'bioenhancement': the introduction of chemicals – such as serotonin – into the brain, or the control of nutrition, sleep, exercise, and stressors that directly alter the beliefs and desires

of individuals (see Persson and Savulescu 2012; Sparrow 2014; Baccarini 2014; Savulescu and Maslen 2015; Clayton and Moles 2018). Toma also alters Diana's interests to bring them into harmony with societal and third party interests. For example, Toma removes certain decisions from Diana, preventing her from driving when her stress levels reach a certain threshold, reporting to her insurance company and government agencies on her behaviour, and from acting in ways that undermine socially cooperative efforts, such as denying options that would increase her carbon footprint. Toma does all of this behind Diana's back; she has no access to the processes by which Toma makes these changes, and they often occur via the 'reading' and altering of subtle patterns regarding Diana's behaviour that she is unaware of (p.27; p.51).

Technologies regulate our behaviours by making certain behaviours possible and constricting others. The regulations that stem from technological artefacts is less obvious than enacted legal norms, and not enacted by a democratic legislator. Its regulative force depends on how engineers, designers and business enterprise bring these artefacts to the market and eventually how consumers or end-users engage with them. Its material and social embedding has a way of inducing or inhibiting certain behaviour patterns, such as sharing personal data. Depending on their design and uptake, technologies can even enforce or rule out certain types of behaviour (p.11).

Toma is of course fictitious, but it is not wildly fantastical. Distributed computing is now a significant feature of our daily existence, and increasingly so. Algorithms are now widely employed – criminal justice and policing, traffic and air-traffic control, finance, environmental monitoring, healthcare, insurance, and education – and their use being extended and diversified. As Hildebrandt has observed, algorithms exhibit a form of *autonomic* decision-making that operates in a manner akin to human decision-making (pp.54-7; see also Hildebrandt 2016). One of the features laid bare by behavioural psychology is that most of our choices are not based on deliberative reasoning, as we may have hoped (p.56; Kahneman 2011, 19-30; Haidt 2012, 32-56). Rather, our decisions are often based on emotional responses, and simple pattern recognition. Our bounded rationality can only cope with a small amount of information, so most functions are performed by our unconscious minds or nervous systems (pp.56-7). If sophisticated reasoning enters the picture at all, then it does so as a form of post hoc rationalisation – an attempt to persuade our self or others of the rightness of our actions (p.56; Haidt 2012, 55). In particular, our moral decision-making is like an elephant and its rider (Haidt 2015). The emotional elephant goes where it wants, and the rider constructs a rationale. Although the rider can, given enough time and effort, alter the course of the elephant, it is rarely in control. If this picture is correct, then the notion of individuals as self-governing, autonomous subjects is a fiction (p.56). Consequently, where algorithmic computing, such as Toma, exercises this influence over our lives, then the line that separates the two becomes blurred. Like Michael Gillhaney in Flann O'Brien's *The Third Policeman*, who has ridden the same bicycle for thirty years, the atoms have become

intertwined, making him almost half bicycle. When not riding, Gilhaney can often be found propped against a wall by his elbow or standing on one leg at the kerb.

The autonomic nervous system does not require our explicit consent to raise blood pressure or increase the breathing speed to adjust our *internal* environment. Similarly, the autonomic computer system could adjust our *external* environment in order to do what it infers to be necessary or desirable for our well-being (p.56).

This blurring of the lines between autonomic systems raise, as Hildebrandt shows, questions about the extent to which humans can be said to be autonomous. One such concern is that when our beliefs, attitudes, and choices are influenced and altered behind our backs, then it undermines notions of freedom and dignity. If it is generally preferable that we hold beliefs because we recognise the force of the reasons behind them, then we should be concerned about non-rational belief formation. When algorithms manipulate beliefs and desires, and the choices that flow from them, it seems like indoctrination. In doing so, it may be thought to fail to treat us with the kind of respect owed to independent human agents. It is not clear where the bicycle ends and the person starts (see also Hildebrandt and Koops 2010, 436).

Types of influence

To unravel what is objectionable, consider an extension of Toma's story.³ Toma judges that Diana's well-being is being hampered by her poor diet and, after reviewing all the available evidence and options, tailored to Diana's physical and psychological composition, concludes that her diet would be much improved by not eating meat. Toma also concludes that changing Diana to a vegetarian diet will have important positive environmental and social impacts that go beyond her own well-being. Diana's earnings are not sufficient for Toma to be able to replace her current poor-quality meat with organically and humanely reared meat, nor is her psychological makeup one that would allow a severe reduction in meat consumption without constantly testing her will-power. The easiest path to better nutrition is for Toma to bring about vegetarianism. Toma is also aware that Diana harbours certain negative emotional reactions to animal suffering and uses this to create leverage. Consequently, Toma: removes meat options from on-line shopping suggestions, replacing them with attractive vegetarian ones; guides Diana toward vegetarian recipes and hides away meat recipes; selects routes, shops, and restaurants that have few or no meat options and well-reviewed vegetarian ones; streams adverts and information to her smartphone about animal suffering; adjusts the image in her smart spectacles to make meat look repulsive; enrolls her onto social media pages of those who oppose animal suffering; administers a small, imperceptible, electrical current via her smart watch that causes nausea whenever Diana's desire for meat threatens to surface. After a short time Diana no longer desires to eat meat. Her well-being improves, and she is pleased to no longer be a meat eater because of the suffering it causes.

³ I thank an anonymous reviewer for suggesting the following example.

Moreover, if presented with the steps taken by Toma, Diana would agree that it is better that Toma took those steps.

What, if anything, is objectionable in this story? One familiar line of argument is that such choosing for others undermines their well-being, either by replacing an individual's judgment about what is in their interests with a less accurate judgment, or in supplanting that judgment it treats that individual a less than fully morally capable being. Both lines of defence face strong objections. Let us consider each in turn.

First, we might think that our choices have *instrumental* value (Scanlon 1998, 251). That is, I am just more likely to judge what is in my interest than some third party, such as the state or its representatives. When I visit a restaurant it is generally better if what appears on my plate is what I selected from the menu, even if the waiter thinks I will enjoy another dish more (Scanlon 1998, 251-2). The waiter is less likely to match my tastes to the options than I am. This account, however, rests on an empirical claim that the increasing power of algorithms has shown to be seriously flawed. What algorithms show is that in some instances I may not be the best judge of my own interests. I may, for example, be swayed by a desire to impress my friends with my knowledge of Greek cuisine and make a choice that I will not enjoy as much as if I had not been swayed by that consideration. Deferring to Toma's authority would prevent me from departing from the rational pursuit of my ends. Algorithms may know us better than we know ourselves.

Second, our choices might have *representative* value (Scanlon 1998, 252): our choices say something about us, even when those choices are less than optimal for our well-being (p.71). When you pick a gift for your mother it matters that it is *you* that picks it (Scanlon 1998, 252). It matters because of what it says about how you view your mother and the relationship between you, even if she would have picked a better gift (one more in line with her preferences) for herself. When Toma makes these choices it removes the opportunity for realising this value.

It is not clear, however, that this account blocks algorithmic enhancement. In fact, algorithms might enhance it by providing more opportunities for the exercise of such choices. If choice has this value, then we have an interest in having more instances of it. Toma is capable of shaping beliefs and behaviour in ways that prevent us from closing off opportunities to make such choices, and of opening up new opportunities. By denying Diana the opportunity to eat meat, it creates a series of future choices with regards to animal welfare, as well as a healthier lifestyle, that she might not otherwise have had.

A further observation is that this account depends upon a controversial view of freedom, one that many reasonable people reject. Many are untroubled by their life plans lacking high levels of choice so long as their ends are the 'right' ones (Quong 2011, 99). It is more important for many that they lead good lives or hold the correct beliefs, and less important how they come to lead it or hold those beliefs. Such independence has a less privileged place in their conceptions of morality. Such individuals might welcome algorithmic enhancement because it makes them more likely to succeed in pursuing what they take to be a good life. Deferring to Toma would be a useful instance of pre-commitment – like Ulysses tying himself to the mast – in order to remove options now that would make the selection of other options at a later point more certain (Elster 1979, 37-47).

If the state can reliably improve the well-being of citizens by utilising algorithms to push them towards better ends, or steer them away from detrimental ends, then there is little in this line of reasoning to block such paternalistic interference. Yet, this seems intuitively implausible. If it were not, then it would commit us to saying there is no difference between Diana coming to act upon reasons that she finds, upon reflection, to be sufficiently weighty to motivate her actions, and acting merely as a result of some unseen external, but benign, force (Clayton and Moles 2018). But, there does seem a difference between Diana rejecting a meat option because she weighs the animal suffering involved as an overwhelming reason not to take a bite, and a feeling of nausea induced by an electric current triggered by her smartwatch. There is something valuable in recognising for oneself the moral requirements that are applicable to us.

Some concerns over influence and self-determination

In this section I lay out some concerns that algorithmic enhancement poses that rely not on their ability to promote well-being, but that operate in spite of this ability. These concerns stem from the idea of self-determination or independence.

It is helpful, following Elster, to distinguish several forms of influence over our beliefs and desires (1979, 81-83). First, a *voluntary choice* is one where a person desires, on the basis of good reasons, x over y , and does x for those reasons. At the opposite end of the continuum is *coercion*. Coercion takes place when a person desires x over y , and continues to do so even when a third party forces her to do y . Between these two poles, other forms of influence exist. *Seduction* is where a person initially prefers x over y , but comes to prefer y over x once she has been coerced into doing y . *Persuasion* occurs when a person who initially prefers x over y is led, by a series of short-term improvements, to prefer y over x . At each step in the process of persuasion the person sees the reasons for change, and comes to accept those reasons *as reasons* for the change.

Diana's vegetarianism looks like a case of seduction: a series of steps that exploit various 'intrapsychic mechanisms' in order to lead her to a desire not to eat meat (Elster 1979, 82). These mechanisms take place behind Diana's back. As such, they differ from persuasion. As Raz writes: 'Manipulation, unlike coercion, does not interfere with a person's options. Instead, it perverts the way the person reaches decisions, forms preferences or adopts goals.' (1986, 377-8). Even if Toma is accurate in terms of the benefits to Diana's health in changing diet, it does not seem to act permissibly. Given the choice, Diana may still refrain from switching. This possibility makes it the case that when Toma acts behind Diana's back, it does so impermissibly. This would not be the case if Diana's friend, Charles, were to attempt to change her mind by arguing with her face-to-face – laying out the various health benefits, environmental impacts, and animal welfare considerations. When Charles engages Diana in such a conversation it is with the explicit intention of changing her mind. Moreover, Diana can argue back, contributing further evidence, or proposing a different relative weighting of the evidence and moral considerations (Elster 1979, 83-4; see also Taylor 1971; Taylor 1976).

It is also worth noting that this form of influence renders Diana's desires unstable. If option a is removed from the choice set (a, b, c) by Toma, and a was

Diana's preferred option, when left with *b* and *c*, Diana might come to view *a* as lacking any real value now it is no longer available (Elster 1983, 112-125). The 'coming to view' element occurs behind Diana's back, perhaps through a psychological process such as the tendency to reduce cognitive dissonance. If this is the case, then it is difficult to determine if any part of the process is autonomous; desires are *adaptive* (Elster 1983, 110-111). When Diana comes to desire vegetarianism via Toma's processes, it is unclear whether she would revert to being desirous of meat eating if Toma were to reverse the process. If Toma were able to cause Diana's desires to flip-flop in this manner, adapting to the available options, then we could be fairly certain that these are not, in any meaningful sense, Diana's actual desires or choices. This would be different from the case where, through learning and experience of, say, her dietary needs or moral concerns about animal suffering, Diana comes to desire vegetarianism.

The acceptability requirement

These concerns provide us with some reason to reject the employment of algorithms where they run the risk of influencing the desires and beliefs of individuals, even where such influences might be thought to improve well-being. Whilst some features of algorithms may enhance our capability to pursue freely chosen ends, when they set what ends we should pursue they undermine our independence. Such manipulation fails to respect the fundamental interest individuals have in the capacity to plan, revise, and rationally pursue their own ends. Such manipulation demeans or diminishes this moral status. Even if governments were certain that some course of action (such as vegetarianism or a specific physical exercise programme) would maximise the well-being of the vast majority of its citizens, implementing a change of belief or desire amongst its citizenry via algorithmic tweaking, would be impermissible because it undermines the respect for independence owed to its citizens. Thus, governments have principled reasons not to take a stand on certain issues regarding the ends that individuals may choose.

This notion of independence also raises issues, as Sparrow (2014, 26) notes, of *political* morality – about the proper relationship of citizen and state. We are born into and inhabit societies governed by a set of legal, socioeconomic, and political institutions and arrangements that coercively impose various kinds of actions upon us. This raises what Rousseau termed the 'fundamental problem' of political society: how to reconcile individual freedom with the constraints necessary to guarantee the security of individuals (Rousseau 1997, 49-50). As free and equal individuals we all have a claim to live under conditions of freedom. But, we also, for our security and prosperity, need to live in societies that are well-ordered – that is, governed by legal constraints (p.10). Part of Rousseau's solution is that our freedom is preserved only if we live under rules that we, ourselves, endorse. When a person endorses the law, we can consider that person as regarding those constraints as self-imposed rules of self-determining, free, individual (Rousseau 1997, 50-1; Rawls 1996, 68).

In Rawls (1996), the view is that citizens are free and equal in virtue of their possession of two moral powers: a capacity for a sense of justice, and a capacity for a conception of the good. A sense of justice is the capacity to understand, apply, and act from a public conception of justice. It supports just political and legal institutions and underpins the proper treatment of others in

accordance with what kinds of behaviour we owe them. A capacity for a conception of the good is the capacity to form, revise, and rationally pursue a conception of one's rational advantage. Rawls states that, in virtue of these two moral powers, persons are free, and that their 'having these powers to the requisite minimum degree to be fully cooperating members of society makes persons equal' (Rawls 1996, 19; see also Hildebrandt 2015, 74-5).

Consequently, we have duties to arrange our institutional framework such that it provides a fair distribution of the various benefits and burdens of social cooperation. This will include the provision of various familiar rights, freedoms and opportunities, and the distribution of socioeconomic goods such that everyone has the means to pursue the ends they choose.

When individuals are free to form their own views, it is inevitable that they will arrive at different judgements about what roles, relationships, or goals are desirable or worthy of pursuit. Even reasonable citizens who are committed to treating each other fairly and with appropriate respect will come to different conclusions about what comprehensive ends are valuable. But, this is *reasonable* disagreement, because each citizen retains the fundamental commitment to the idea of treating each other as free and equal, as well as to an ideal of social unity where citizens see themselves as 'ready to propose fair terms of cooperation and to abide by them provided others do' (Rawls 1996 54, 63-66; Quong 2011, 291).

What follows from this is that the state should be guided by a conception of political morality that is acceptable to free and equal citizens. This 'acceptability requirement' claims that laws and policies lack justification to the extent that citizens can reasonably reject the moral ideals and principles that guide it (Clayton and Stevens 2018). If we have reasons to arrange our institutions in ways that preserve or maintain independence, then when the state appeals, in justification of a law or policy, to the worth of any particular comprehensive end, then its reasons are likely to be rejected by those citizens who do not share that conception of the good. As Hildebrandt argues, this does not result in a view that eschews all constraints on freedom, but only *unreasonable* constraints (p.80). Reasonable constraints are those that are acceptable to reasonable citizens, given the myriad of convictions about comprehensive ends that are an inevitable outcome of the exercise of practical reason under free, democratic institutions. In other words, the justification of the state's powers to coerce citizens or to prevent certain actions must proceed in terms that do not deny the assumptions, ideals, or conclusions of the diversity of doctrines held by reasonable citizens, that is, citizens who respect the rights and interests of others.

If the state were to mandate the use of algorithms to interfere in society in order to promote the well-being of its citizens behind their backs, then it would be at odds with the acceptability requirement. It would be at odds with it because it would be seeking to promote ways of living or conceptions of what constitutes human flourishing by cultivating certain comprehensive beliefs and shaping the choices of individuals in accordance with those ends. Such conceptions would be controversial and subject to rejection by some reasonable citizens. Consider the following examples of governments using algorithms to manipulate the desires of citizens to improve their well-being or to bring about socially beneficial ends: First, the de-emphasising of consumerism over eco-friendly lifestyles by making some options (pubic transport, recycling) more

visible, or by altering the physical or psychological environment to make the formation of beliefs and desires contrary to the state-mandated ones less likely. Second, the promotion of certain diets via the mechanisms considered above with the aim of improving health and for reducing environmental impacts. Third, the promotion of a Christian lifestyle and beliefs because it is considered not only true but that a shared set of religious values is socially beneficial for achieving peace and stability.

If the above seem far-fetched, then we need only recall how extremely simple adjustments to both our external environments and our psychological mechanisms can have significant impact on our beliefs, desires, and behaviour. For example, bitter tastes can trigger moral disgust and sweet tastes can trigger more favourable moral judgments (Eskine et al 2011). Physical touch can trigger higher levels of trust and monetary sacrifice between strangers (Morhenn et al 2008). Priming individuals with favourable images and stories of individuals from different cultural backgrounds mitigates discriminatory tendencies such as anti-immigrant prejudices (Motyl et al 2011). Dopamine levels and the number of friends an individual has are, in combination, correlated to political views (Settle et al 2010). The use of emotive words can lead to the same moral statements being judged differently (Van Berkum et al 2009). Poorly lit environments lead to increases in cheating and self-interested behaviour (Zhong et al 2010a). Physical cleanliness leads to more severe moral judgments immediately after the process of cleaning (Zhong et al 2010b). All of these are easily manipulated by smart technology. Yet, in taking any of these actions, the state seemingly endorses views that, though they might be conducive to either individual well-being or socially beneficial ends, rely on controversial claims, such as the causes and relative weighting of reasons for climate change and the processes necessary for combating it, the importance of animal welfare and the moral necessity of vegetarianism, or the plausibility of a religious doctrine and the appropriateness of its values for social cohesion.

Cultivating a sense of justice

The acceptability requirement is a stringent test for laws and policies to pass, and it rules out many possible uses of technology, such as policies aimed at promoting the well-being of citizens. It does not, however, rule out all forms of moral enhancement. I argue in this section that the use of such technology to cultivate a sense of justice in citizens is morally permissible.

A sense of justice is the capacity to understand, apply, and act from a public conception of justice. It supports just political and legal institutions and underpins the proper treatment of others in accordance with what kinds of behaviour they are owed by us. A sense of justice is necessary if we are to lead independent lives. Without the appropriate limits on our behaviour that a sense of justice brings, the pursuit of our own ends would undermine the independence of others. A sense of justice does not set the whole content of what is owed to others, it only sets limits on what is permissible in terms of actions that would make it more difficult for others to realise their own interests.

To see this, we can note that the acceptability requirement does not, and cannot, apply to all moral rules. For example, where there are rules or laws prohibiting the deliberate harming of others, it is irrelevant whether an individual to whom they apply endorses them or not. That is, there are morally

enforceable duties – to do or refrain from doing certain things – that are not subject to reasonable agreement. Whilst it may be valuable for individuals to endorse such rules, their non-acceptance does not provide grounds for thinking those rules lack validity, as all the rules do is ensure that the individual does what she is morally required to do. Only in those cases where an individual is not under an enforceable moral duty not to act in a certain way or another is that individual's freedom violated if she is subject to laws she reasonably rejects.

Belief formation plays a vital role in helping citizens to develop an effective sense of justice and a conception of citizenship. A form of education that cultivates such attitudes is not ruled out by the acceptability requirement. Even those, such as children, who cannot consent are reasonably subject to a state-mandated curriculum that attempts to cultivate such beliefs and attitudes. As Rawls writes, the role of education will include:

[A] knowledge of their constitutional and civic rights so that, for example, they know that liberty of conscience exists in their society and apostasy is not a legal crime, all this is to ensure that their continued membership when they come of age is not based simply on ignorance of their basic rights or fear of punishment for offenses that do not exist. Moreover, their education should also prepare them to be fully cooperating members of a society and enable them to be self-supporting; it should also encourage the political virtues so that they want to honor the fair terms of social cooperation in their relations with the rest of society (Rawls 1996, 199).

Part of this function of education is *directive*: to bring individuals to a set of beliefs about the value of social unity and to instil attitudes about the proper treatment of others. Cultivating this sense of justice is part of the role of education, and it does so, often at a young age, behind the backs of individuals via the use of non-rational mechanisms such as rewards and punishments, issuing prescriptions, and modelling the behaviour of others. However, education for these ends has had limited success. Some people fail to cultivate certain beliefs, desires, and forms of behaviour to a sufficient degree. Where this is the case, the state's use of technological enhancement to better secure the uptake of these beliefs, attitudes, and norms of behaviour may be permissible. The kinds of autonomic computing systems that power Toma might be brought to bear on such matters. Using the same range of options, algorithms might be employed to combat latent discriminatory tendencies, on grounds such as race, gender, sexuality, or age, and to pre-empt desires to deliberately harm others or conduct gratuitous violence. By filtering information such that the user is less likely to encounter, say, extreme racist or homophobic literature, images, or video, or by managing meals, diet, and surrounding temperatures, noises, and sleep patterns, some influence can be exercised over the resulting attitudes.

Two objections

The above argument raises two immediate criticisms, the exploration of which allows for the view defended to be fleshed out in more detail, although the observations will be necessarily brief.

The first objects that the asymmetric treatment conceptions of the good and a sense of justice is not sustainable on principled grounds. If the state may not impose, or appeal to, particular comprehensive ends in its justification of its policies, because these ends are subject to reasonable disagreement, the criticism runs, then similar disagreement exists with regards to what justice requires. After all, individuals do not simply disagree about what makes their lives go well, they disagree about almost every value, including freedom, equality, and justice itself. Political debate about matters of justice, such as taxation, healthcare, and social security, seem as intractable and deeply motivated as debates about abortion, fox hunting, or prayer in schools. This suggests there are no shared reasons or uncontroversial views that would form the foundation for state policies about what justice requires. When the state, say, enacts policies to reduce economic inequality, some citizens will be subject to laws and constraints they think lack validity in just the same manner as if the state had implemented a Christian view of morality on atheistic citizens.

This presents a dilemma. Either we admit that reasonable disagreement applies to matters of justice and retract a partisan commitment to cultivating a sense of justice amongst citizens; or, we deny that questions of justice are subject to reasonable disagreement, and extend this to matters of the good. Either horn of the dilemma would undo the view defended above. Grasping the first horn would make the use of algorithms to supplement moral learning impermissible. Grasping the second would commit us to a perfectionist account of political morality that I have been at pains to avoid because it would permit the use of algorithmic enhancement to advance particular state-mandated versions of human flourishing.

The response to this objection is that it ignores the fact that the acceptability requirement appeals only to *reasonable* views, not to views tout court. A reasonable view, on Rawls's conception, is marked out by its being willing to propose and abide by fair terms of social cooperation, and to view citizens as free and equal. A reasonable conception of justice is, in turn, marked out by several criteria, including the provision and protection of a familiar set of rights and liberties, certain opportunities, a specific division of wealth and income, universal healthcare, and the state being employer of last resort. These goods are necessary in order for individuals to pursue their self-chosen ends and to cultivate and act from a sense of justice. This means that some views – such as Nozick's libertarian view – would be excluded as unreasonable because they deny certain all-purpose resources as necessarily accompanying liberties in order that individuals can make full use of those liberties. Nozick's view claims, instead, that the unfettered market is the appropriate mechanism for setting the baseline distribution of such resources.

The response to the asymmetry objection, then, is not to deny the existence of disagreement about justice, but to point to the fact that much actual disagreement is unreasonable, and thereby irrelevant to the permissibility of the state deploying its coercive power. Reasonable disagreement will still occur over justice – for example, Rawls's own conception, 'justice as fairness', is but one reasonable conception amongst a range of reasonable conception, but the scope of the disagreement will be narrower. Consequently, there would still be some latitude for argument and debate over what the content of a sense of justice could be promoted through algorithmic enhancement.

The second objection is that in seeking to bring about beliefs and desires conducive to a sense of justice or social unity, the state undermines its own legitimacy. It does so, this objection claims, because the sense in which an individual can be said to freely consent to the authority of the state is weakened when the state itself shapes the political motivations of its citizens. If beliefs and desires are shaped in this way, then we cannot, as we have seen, be certain that they are genuine, rather than adaptive. It would be unsurprising if citizens whose political motivations are shaped by the state come to endorse the view of justice on which the state is based. This is a criticism brought against education for a sense of justice in general by Brighouse (1998). To ward against this Brighouse claims that such endorsement must: (1) be based on political arrangements which citizens would consent to if those citizens were reasonable, possessed sufficient information, and capable of reasoning in a manner. However, Brighouse worries that this hypothetical consent is too easily attained, so he adds: (2) there must be sufficient *actual* consent, and this too must be free, informed, and rational (Brighouse 1998, 723).

In response I offer two brief observations. First, actual consent is too high a bar. Brighouse claims that actual consent is the 'usual' criterion for a theory of political morality to meet. But, this is not the case if he means that the most plausible accounts do, or must, include this requirement. The views of Rawls, Raz, and Dworkin (to name some of the most prominent) do not consider consent to be a requirement of a view's legitimacy. To require this would be to impose a threshold that no view could meet. There are always those – nihilists, free-riders, the infirm, and the bloody-minded – who would withhold consent. Rawls, for example, relies upon a 'natural duty of justice' to generate legitimacy (Rawls 1971, 114-117). Dworkin adopts an account of associative duties to ground consent (1986). And, Raz employs a mechanism to generate an obligation to obey the law based on the idea that consent exists where compliance with a given law thereby enables a person to better act upon her reasons for action than not complying with it (1988).

Second, the claim that hypothetical consent is too easily obtained is also open to challenge. A plausible view of political morality rests on stringent ideals of justice and democracy. These will place significant demands on political principles and institutions, not ones that are weak and easy to meet (Clayton 2006, 134). Rawls's natural duty of justice is a case in point. The duty to bring about, support, and comply with just institutions and laws is a significant requirement that places considerable demands on citizens, including the duty to obey laws because they are laws brought about by just institutions. That such features must be present for the state to be legitimate imposes a high threshold on what would count as hypothetical consent.

Because of this high threshold we can see why many of the problems in advancing justice are collective action problems that require the coordination of all citizens and the assurance that the benefits and burdens of this cooperation are distributed fairly. The natural duty of justice generates an obligation to obey the law when this better enables us to conform with the requirements of justice than if we each, individually, acted according to our own judgments about how to realise justice. Because of this, cultivating a sense of justice supportive of just institutions is crucial. Moreover, this must be true for a sufficient number of citizens. If the number of citizens who possess a sense of justice is too small, then

it becomes counter-productive: those who act from a sense of justice would be taken advantage of by those who would free-ride. Whilst we should prefer methods, such as formal education, where individuals come to adopt a sense of justice based on an understanding of the reasons, if a sufficient number of citizens cannot be achieved in this way, then some supplementary options such as algorithmic enhancement are, at least in principle, permissible.

Conclusion

Despite familiar worries, algorithms offer significant opportunities for enhancing human well-being. They can provide ways of improving decision-making, selecting, and advancing valuable goals. They offer ways of enabling governments to better advance the well-being of their citizens by providing alternatives to legal constraints and incentives. In this contribution I have argued that, despite such purported benefits, there are principled reasons for governments to refrain from employing technology to these ends. Governments, I have argued, cannot promote comprehensive goals or ends without undermining their own legitimacy. Instead, laws and policies must be acceptable to citizens who will, necessarily, disagree about the value or importance of these comprehensive ends. I have also argued, however, that this does not apply to the promotion of a sense of justice. In order to maintain just and stable political institutions it is necessary that a sufficient number of citizens come to adopt and abide by the appropriate attitudes and forms of behaviour. Where traditional methods of belief formation, such as educational institutions and informal methods of socialisation, fall short, then algorithmic enhancement can permissibly pick up some of the slack.

References

- Baccarini, E. (2014) 'Public Reason and Moral Bioenhancement'. *Ethics and Politics*. Vol XVI, 2, 1027-1041.
- Brighouse, H. (1998) 'Civic Education and Liberal Legitimacy'. *Ethics* 108. 719-45.
- Clayton, M. (2006). *Justice and Legitimacy in Upbringing*. Oxford: Oxford University Press.
- Clayton, M., and A. Moles (2018) 'Neurointerventions, Morality and Children' in D. Birks and T. Douglas (eds.), *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford: Oxford University Press. 235-251.
- Clayton, M., and D. Stevens (2018) 'What's the Point of Religious Education?' *Theory and Research in Education* 16. 65-81.
- Elster, J. (1979) *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

- Elster, J. (1983) *Sour Grapes*. Cambridge: Cambridge University Press.
- Eskine, K., N. Kacinik and J. Prinz (2011) 'A Bad Taste in the Mouth: Gustatory Disgust Influences Moral Judgment'. *Psychological Science*, 22, 3. 295-299.
- Haidt, J. (2012) *The Righteous Mind: Why Good People are Divided by Politics and Religion*. London: Penguin.
- Hausman, D., and B. Welch (2010). 'Debate: To Nudge or Not to Nudge?' *The Journal of Political Philosophy* 18, 1. 123-136.
- Hildebrandt, M. (2013) 'Slaves to Big Data. Or Are We?' IDP. REVISTA DE INTERNET, DERECHO Y POLÍTICA, 17, 7-44. Available at: <https://www.raco.cat/index.php/IDP/article/viewFile/303366/393038>
- Hildebrandt, M. (2015) *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*. Cheltenham: Edward Elgar Publishing.
- Hildebrandt (2016) 'The New Imbrogio: Living with Machine Algorithms' in L. Janssens (ed.) *The Art of Ethics in the Information Society*. Amsterdam: Amsterdam University Press. 55-60.
- Hildebrandt, M., and B.-J. Koops (2010) 'The Challenges of Ambient Law and Legal Protection in the Profiling Era'. *The Modern Law Review*, 73, 3. 428-460.
- Kahneman, D. (2011) *Thinking, Fast and Slow*. London: Penguin.
- Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., and L. Floridi (2016) 'The ethics of algorithms: Mapping the debate'. *Big Data and Society*. July-September. 1-21.
- Morhenn, V., J. Park, E. Piper, and P. Zak (2008) 'Monetary Sacrifice Among Strangers is Mediated by Endogenous Oxytocin Release after Physical Contact'. *Evolution and Human Behavior* 29. 375-383.
- Motyl, M., J. Hart, T. Pyszczynski, D. Weise, M. Maxfield, and A. Siedel (2011) 'Subtle Priming of Shared Human Experiences Eliminates Threat-Induced Negativity Towards Arabs, Immigrants, and Peace-Making'. *Journal of Experimental Psychology* 47. 1179-1184.
- Pasquale, F. (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. London: Harvard University Press.
- Persson, I., and J. Savulescu (2012) *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press.
- Quong, J. (2011) *Liberalism Without Perfection*. Oxford: Oxford University Press.
- Rawls, J. (1971) *A Theory of Justice*. Oxford: Oxford University Press.

- Rawls, J. (1996) *Political Liberalism*. New York: Columbia University Press.
- Rawls, J. (2001) *Justice as Fairness: A Restatement*. Cambridge, Mass.: Harvard University Press.
- Raz, J. (1986) *The Morality of Freedom*. Oxford: Clarendon Press.
- Rosa, H. (2013) *Social Acceleration: A New Theory of Modernity*. New York: Columbia University Press.
- Savulescu, J., and H. Maslen (2015) 'Moral Enhancement and Artificial Intelligence: Moral AI?' In Romportl, J., Zackova, E., and J. Kelemen (eds) *Beyond Artificial Intelligence: Topics in Intelligent Engineering and Informatics*, vol. 9. Springer, Cham. Pp.79-95.
- Scanlon, T. (1998) *What We Owe to Each Other*. Cambridge, Mass.: Belknap Press.
- Settle, J., C.Dawes, N.Christakis and J.Fowler (2010) 'Friendships Moderate an Association between a Dopamine Gene Variant and Political Ideology'. *The Journal of Politics* 72, 4. 1189-1198.
- Sparrow, R. (2014) 'Better Living Through Chemistry? A Reply to Savulescu and Persson on "Moral Enhancement"'. *Journal of Applied Philosophy* 31, 1.
- Sunstein, C. (2014) *Why Nudge? The Politics of Libertarian Paternalism*. London: Yale University Press.
- Taylor, C. (1971) 'Interpretation and the Sciences of Man', *The Review of Metaphysics*, 25,1. 3-51.
- Taylor, C. (1976) 'Responsibilities for Self' in A.O.Rorty (ed.) *The Identities of Persons*. Berkeley, CA.: University of California Press. 281-299.
- Van Berkum, J., B.Holleman, M.Mieuwland, M.Otten and J.Murre (2009) 'Right or Wrong? The Brain's Fast Response to Morally Objectionable Statements'. *Psychological Science* 20, 9. 1092-1099.
- Wilkinson, T.M., (2013) 'Nudging and Manipulation'. *Political Studies* 61. 341-355.
- Zhong, C-B., V.Bohns, and F.Gino (2010a) 'Good Lamps Are the Best Police: Darkness Increases Dishonesty and Self-Interested Behavior'. *Psychological Science* 21, 3. 311-314.
- Zhong, C-B., B. Strejcek, and N.Sivanathan (2010b) 'A Clean Self Can Render Harsh Moral Judgment'. *Journal of Experimental Social Psychology* 46. 859-862.

Index terms

Autonomic
Autonomy
Big Data
Choice
Coercion
Conception of the good
Consent
Education
Framing
Justice

- Justice as fairness
- Natural duty of

Legitimacy
Machine Learning
Manipulation
Moral enhancement
Nudging
Persuasion
Pre-commitment
Preferences

- Adaptive

Reasonable
Reasonableness
Seduction
Toma
Ulysses