Note: this is an article whose final and definitive version is published in the *Journal of Medical Ethics*; the official version of this article is available online via: http://jme.bmj.com/content/early/2016/08/24/medethics-2016-103593.short.

Climate change, cooperation, and moral bioenhancement

Toby Handfield, Pei-hua Huang, and Robert Mark Simpson

Abstract**:** The human faculty of moral judgment is not well suited to address problems, like climate change, that are global in scope and remote in time. Advocates of 'moral bioenhancement' have proposed that we should investigate the use of medical technologies to make human beings more trusting and altruistic, and hence more willing to cooperate in efforts to mitigate the impacts of climate change. We survey recent accounts of the proximate and ultimate causes of human cooperation in order to assess the prospects for bioenhancement. We identify a number of issues that are likely to be significant obstacles to effective bioenhancement, as well as areas for future research.

1. Introduction

Many factors make it difficult for humans to forge the collective commitments that are necessary to mitigate the effects of anthropogenic climate change. The data that underpin warnings about the dangers are complex and hard for non-experts to understand [1]. Many of the dangers are remote in time and therefore easy to ignore [2]. Human beings are sometimes prone to wishful thinking and undue optimism, especially in the face of uncertainty [3]. And we are also inclined to prioritise the interests of our immediate kin over the needs of faceless others in the future [4].

Optimal responses to the risks of climate change will almost certainly require some parties to make sacrifices now in order to help others later [5,6]. Humans typically exhibit some altruistic inclinations, but there are reasons to expect that this sort of altruism will be of limited value in responding to problems on the scale of climate change.

Recent advocates of 'moral bioenhancement' have argued that societies should make pharmacological and/or genetic interventions to boost people's altruistic dispositions [7-9]. They claim this will greatly improve our chances of effectively addressing climate change, proliferation of nuclear weapons, and related global challenges [7]. According to two prominent advocates of moral bioenhancement, Ingmar Persson and Julian Savulescu, there are four major dispositions involved in our moral cognition that could be beneficially influenced via pharmacological or genetic therapies in the future, namely: (i) the disposition towards altruism, (ii) the disposition towards fairness, (iii) the tendency to adopt a causal conception of responsibility (which leads us to focus on harms that we cause, at the expense of neglecting possible benefits that we fail to confer), and (iv) the short-term bias in our decision-making. The first two dispositions could be enhanced, they say, whereas the latter two could be inhibited. While Persson and Savulescu concede that proposals along these lines can only be speculative as things currently stand, they are optimistic that productive interventions to increase altruism and empathy could be brought about in the future. And given the magnitude of the problems that such moral bioenhancement could help to address, they recommend this as a priority area for research investment [7].

A number of ethical concerns have been raised against moral bioenhancement. Political programs in which governments (or other bodies with coercive powers) seek to modify people's thoughts and feelings to try to fix social problems may be intrinsically objectionable [10,11]. And there might be reasons to worry what kind of moral personhood would survive in the wake of moral bioenhancement, were it to come to fruition [12-14].

Here we set aside these ethical concerns about moral bioenhancement, and instead review relevant evidence and models to assess the gains that are likely to result from intervening on moral dispositions. We argue that enhancing these dispositions could in fact hinder the kind of cooperative efforts that are required for climate change mitigation. The problem, in essence, is that our moral dispositions operate in a strategic environment that contains both opportunities and threats. While there are likely to be opportunities to bring about more fruitful cooperation by enhancing some psychological traits, doing so will simultaneously leave those treated more susceptible to deception and exploitation. This in turn makes it easier for harmful or antisocial behaviour to carry on unimpeded. Amplifying other dispositions, such as the disposition to monitor and sanction transgressions from a cooperative scheme, is unlikely to remedy this problem without introducing further difficulties.

Despite the fact that there is a significant and fast-growing scholarly discourse on the ethics of moral enhancement, one might think that the kind of proposals put forward in this literature are too

speculative to be taken seriously as a subject of scientific inquiry. We do not currently possess the technological means to make any kind of reliable or precise adjustments to moral dispositions like trust and empathy, nor do we have institutional mechanisms that would facilitate a widespread implementation of such technologies if they were to be realised any time in the near future. But rudimentary moral bioenhancement techniques are already available (such as the oxytocin treatments we discuss below; see also [15]), and bioenhancement programs of a certain sort are already coercively implemented in some jurisdictions, e.g. where chemical castration is involved in the sentencing of people convicted of particular types of criminal acts [16].

2. The evolution of cooperation

Proponents of moral bioenhancement say that we should increase or amplify (some of) our prosocial dispositions. From a biological perspective, however, the degree of prosociality we see in humans is already surprisingly high. Some kind of explanation is needed to account for the prevalence of prosocial and cooperative dispositions in humans, given that such dispositions seem to make individuals more vulnerable to being hurt or exploited, and hence at a competitive disadvantage to selfish others.

2.1 Altruistic cooperation

The prisoners' dilemma is a comprehensively studied paradigm for modelling altruistic social interactions. In the most general form of a prisoners' dilemma, individual players can offer help at cost c, thereby conferring benefit $b > c$ on the other player. Each individual's payoff is maximized by not helping, but the net result of mutual helping (each player receives $b - c$) is greater than if neither player helps (both players receive 0). Consequently, if players pursue their individual advantage, they will bring about a socially suboptimal outcome. These payoffs are represented using nominal values ($b = 4$, $c = 1$) in Figure 1. (Hereafter we follow the custom of referring to the more prosocial behaviour of helping in this, and related social dilemma games, as "cooperating", and refer to not helping as "defecting".) This is regarded as a model of potentially *altruistic* behaviour because each player can benefit the other, but only at personal cost. The best possible outcome is to benefit from someone else's cooperation, while choosing to defect. (We discuss *mutualistic* cooperation – where both parties do best when both cooperate, in the following section.)

	Cooperate	Defect
Cooperate	3,3	−1,4
Defect	4,−1	0,0

Figure 1. Prisoner's dilemma payoff matrix

It is well established that in controlled experimental versions of the prisoners' dilemma people are willing to at least initiate cooperation – contrary to apparent self-interest – at a nontrivial rate [17]. This occurs despite the prediction that, absent other factors, rational players would not cooperate. The choice to defect dominates the choice to cooperate; which is to say, irrespective of the other player's choice, each player will fare better by defecting instead of cooperating.

To explain how organisms can evolve to reliably cooperate in situations of this type, there are two broad classes of mechanism that have been postulated. One class involves introducing factors that structure the population, making the interactions that occur non-random. If the population is structured in such a way that cooperators interact with other cooperators at a greater than chance frequency, it is possible to sustain non-zero levels of cooperation in an evolutionarily stable population. Mechanisms such as kin selection (helping genetic relatives) [18], direct reciprocity ("I'll help you if you help me") [19], and indirect reciprocity ("I'll help you if I see you have helped others") [20] all work in this fashion. The second class of mechanism involves changing the payoffs associated with defection and cooperation, for instance, by introducing sanctions that are imposed on defectors. This is exemplified by models of strong reciprocity ("I'll punish you if I see you have failed to help others") [21].

The following psychological traits are all likely to be involved in implementing the broad evolutionary mechanisms identified above.

> *Parochialism* – the benefits of reciprocity generally require that groups are not too large and that we have a reasonable probability of interacting again with those we have successfully interacted with in the past. Some degree of in-group bias is likely to be an important element in maintaining the viability of at least some altruistic behaviours, because interactions with outgroup members are likely to involve fewer repeat encounters and are likely to be undertaken with less information about past behaviours [22-24].

Reputation monitoring – we pay close attention to the cooperative behaviour of others, and make our future cooperative efforts conditional upon what we know about the behaviour of others [25-27].

Retribution – at least some of us may be disposed to retaliate against those who betray our trust or abuse our generosity [28-31].

Parochialism and reputation monitoring contribute to structuring the population so as to implement mechanisms such as kin selection, direct reciprocity, and indirect reciprocity. Retribution contributes to mechanisms that involve changing the payoffs of defection, such as strong reciprocity. All three dispositions are in some sense negative or defensive: they do not directly lead agents to choose behaviours that confer benefits on unrelated others, and may motivate mutually costly behaviour, such as punishment. In the presence of these dispositions, however, overtly altruistic dispositions such as a tendency to trust others or to empathise with others can be adaptive.

Empathy/kindness – if we are moved by the plight of others, it will be aversive for us to see them suffer, so we are more likely to provide them with assistance [32-34]. This is what we mean by empathy or kindness: a disposition to regard the conferral of benefits on others as inherently desirable.

Trust – some prosocial behaviour requires making oneself vulnerable to being betrayed, exploited, or let-down by others. In this context, *trust* is a willingness to make oneself vulnerable in this way, for the sake of a cooperative or altruistic goal. This trait is distinct from generalised attitude toward risk [35,36].

Without denying the reality of these traits, it is a matter of common sense that they operate in limited, conditional, and context-sensitive ways. This accords with our theoretical understanding of the fragility of altruistic behaviour in evolutionary contexts. All favoured models entail that such altruistic dispositions will be selected for only where they operate in conjunction with mechanisms to guard against exploitation, such as the parochial, retributive, and judgemental dispositions described above. No credible account has been given of how indiscriminate empathy and trust could, in isolation, be favoured by selection pressures. If a number of organisms began to display indiscriminate empathy and trust in strategic environments like this, it would amount to playing a dominated strategy: those who lacked the novel traits would prosper, and the mechanisms of selection would act to extinguish trust and empathy from the population.

Three further pieces of evidence suggest that the disposition to trust, in particular, cannot be unilaterally enhanced without destabilising a prior equilibrium involving defense against potential exploitation.

(a) *Economic experiments* employing the trust game, a paradigm designed to test participants' willingness to cooperate together to gain better rewards, have consistently found that the return to trusting behaviour is approximately the same as the return to non-trusting behaviour, despite substantial variation in the level of trust shown between different cultures [37]. The game involves two players: an "investor" and a "trustee". Investor may transfer any portion of her endowment to the trustee's account. The amount transferred is multiplied by a rate of return >1. Trustee can then choose to transfer any amount of the multiplied quantity back to the investor, and keeps the remainder. If the trust shown by the investor is reciprocated, both parties benefit. Typical investors transfer roughly half of their initial endowment, and a non-trivial proportion of trustees return at least that much to the investor [37]. Also typically, a number of trustees fail to reciprocate trust by returning zero or other amounts less than the initial investment, but subjects appear to accept a social norm requiring that the profits be shared with the investor [38].

In an experiment conducted in Zurich, investors make accurate discriminations of degree of trustworthiness between residents of different urban districts, and invest more in regions that yield higher average returns [39]. This evidence does not directly support a causal inference as to what would happen if some individuals unilaterally increased their degree of trust, but it is suggestive that subjects are adjusting decisions to trust in a way that is sensitive to the strategic environment. This makes it unlikely that there are large gains to be obtained by intervening to modify degree of trusting behaviour at the population level.

(b) *Recent studies on oxytocin* show that there may be some linkage between the propensity to trust and parochial tendencies to guard against out-group members. Oxytocin is a naturally-produced neurotransmitter which affects people's dispositions toward generosity [40] and trust [41,42]. In one study using the trust game, subjects who had an oxytocin nasal spray administered to them were significantly more likely than control group participants to opt for a maximally trusting choice, by investing all of their money [41]. Prima facie, such findings support the possibility of effective moral bioenhancement, since they show that pharmacological interventions can promote trust among strangers.

However oxytocin has also been found to promote in-group bias and parochialism [43,44]. One study found that in an implicit association test, subjects treated with oxytocin were, compared to control group subjects, faster to associate negative phrases with names linked to an ethnic out-group, and faster to associate positive phrases with names linked to an ethnic in-group [44]. Another found that subjects treated with oxytocin were more likely to make financial decisions that were adverse to out-groups in settings where the in-group was exposed to risk of loss [43]. These observations are consistent with the idea

that trust is an inherently risky disposition, and cannot evolve without accompanying traits that protect against exploitation.

(c) *Social attitude surveys* find that the distribution of income is non-monotonic and hump-shaped with respect to trusting attitudes [45]. The highest income individuals have median levels of trust relative to the local population, and the most trusting individuals earn approximately 14% less than those with median trust levels. Furthermore, highly trusting individuals report higher rates of having been cheated in the past, lending credence to the hypothesis that their lower income is a result of more trusting individuals being exploited [45].

Now consider again the social problems associated with climate change – problems whose solution would require a collective commitment by a large and diverse group of parties to sacrifice some of their economic interests in the immediate term (e.g. by dramatically reducing global CO_2 production) so as to limit the consequences that future generations will have to cope with. If we could be confident that interventions to increase prosocial dispositions like trust and empathy would lead to a social order governed by complete and perfect trust – a social order in which individuals always cooperate for the sake of mutual benefit, and in which there is no-one who betrays or exploits others for personal gain – then we would have some reason to favour such interventions. But interventions like these are absurdly utopian. More realistically, we can expect interventions that significantly increase the elementary prosocial dispositions of a significant number of people. But in view of the strategic implications of trusting others, increased elementary prosocial dispositions will not translate into a reliably increased rate of cooperation. It will instead be an environment that is congenial for infiltration by exploitative, uncooperative agents.

2.2 Mutualist cooperation

Our discussion to this point has assumed that the mode of prosocial behaviour that is relevant to social problems connected with climate change is something like cooperation in a prisoners' dilemma. This is not an idiosyncratic assumption; many problems linked to climate change are naturally construed as tragedies of the commons [46], and the game theoretic form of this dynamic is routinely modelled using the prisoner's dilemma. Still, this way of framing the problem may be too restrictive, and it is possible that other modes of prosocial cooperation are important in addressing problems of climate change. Mutualistic cooperation is arguably central to the evolution of human cooperation in general [47,48], and it provides another framework via which we can model what is required to sustain effective cooperation in

complex social situations. The stag hunt is the standard game theoretic tool for modelling mutualistic co-operation. If both players work together to hunt a stag (mutual cooperation), they each receive a high payoff (4,4); but hunting stag is risky, and if one player abandons the hunt to catch a hare instead (one cooperates, the other defects), then the stag hunter will go hungry while the hare hunter at least gets something (0,3). Both situations in which the players adopt an identical strategy – stag and stag (mutual cooperation) or hare and hare (mutual defection) – are Nash equilibria, i.e. scenarios in which no player stands to gain by unilaterally altering her strategy. Obviously, however, it is better – for both players individually, and for the good of the social group that they belong to – if players can be induced to cooperate in hunting stag, so as to achieve the more rewarding equilibrium.

Bioenhancement of either a disposition to trust or of a tendency to empathise can be predicted to lead to a higher rate of cooperation in a stag hunt (see Figure 2). Both factors will increase the relative desirability of cooperating. Enhanced trust will lead to a greater tolerance of the risk associated with co-operation. Enhanced empathy will increase the aversiveness of making a choice that could lower the pay-off for the other party.

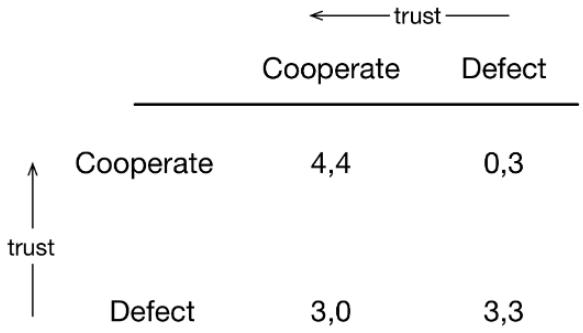

Figure 2. The effect of enhancing trust in a stag hunt. The game has two equilibria: a risk dominant equilibrium (D,D) and a payoff dominant equilibrium (C,C). Trust increases willingness of players to make the high risk choice, thereby moving them to the high-payoff equilibrium. Enhanced levels of empathy would have a similar effect, by making the CC payoff relatively more attractive than all other payoffs.

Agreeing to major collective reductions in CO_2 emissions in order to limit atmospheric warming is a more complex strategic situation than a schematically modelled stag hunt, but it could share a similar underlying structure [49]. If a large enough proportion of parties make significant cuts, we will all achieve benefits. If only a few parties make significant cuts, then those who cut will still stuffer the negative consequences of

climate change, but they will also incur the additional burden of having forgone short-term benefits arising from the exploitation of fossil fuel resources. It is strongly desirable to achieve the better equilibrium, but this requires a high degree of trust between the parties.

However, although increasing trust will be beneficial in mutualistic encounters of this sort, it remains doubtful that an elevation of our tendency to trust others will bring mutual benefits across the whole range of human interactions. The real world is rife with strategic uncertainty: the payoffs of our various courses of action are difficult to discern, even after the fact; and we can expect other agents to react to our new decisions in ways that change those payoffs further. If we demonstrate an increased willingness to trust, we should anticipate an increase in the frequency of agents who devise schemes to exploit that trust.

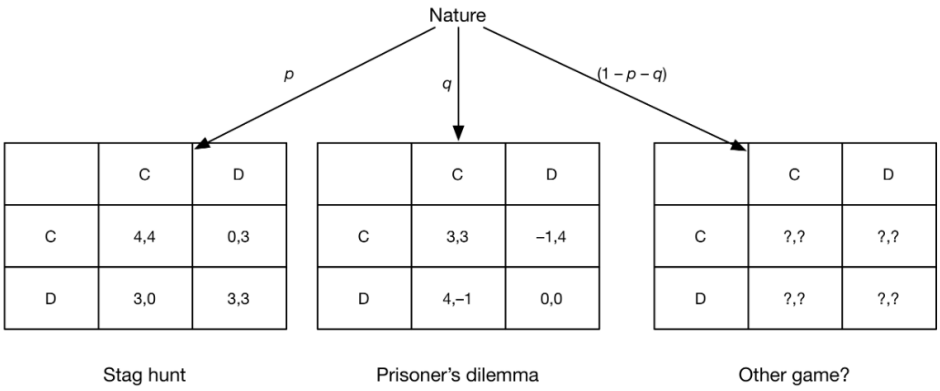

Figure 3. A Bayesian game against nature, in which an agent must estimate the probability that she faces a game involving mutualistic cooperation, altruistic cooperation, or other possibilities. Provided p is high, trusting behaviour will be adaptive, but if agents with high estimates of p become frequent, it is likely that exploitative strategies will evolve and thereby increase the frequency of encounters that lead to the exploitation of trust.

A better model of the situation, then, is one where an agent faces a Bayesian game against nature (see Figure 3). Nature determines whether we are playing a prisoners' dilemma, a stag hunt, a coordination game, or some other sort of social interaction. An agent has some estimate of the probability distribution over these possibilities, and attempts to make an optimal decision in the face of this uncertainty. An agent who is more disposed to trust will be an agent who is more disposed to accept that the game she faces is a cooperative one, more disposed to think that her partner will make a cooperative move, or both. For these reasons, she will be more likely to cooperate.

An agent who is trusting in these ways, however, will thereby create an environment favourable to strategies which exploit the greater degree of trust. Evolutionary reasoning predicts that exploitative

agents will increase in number, and so the frequency with which the agent encounters stag hunts will decrease. The trusting agent will be faced more frequently with games where cooperative behaviour is maladaptive. In short, it is not possible to unilaterally enhance prosocial dispositions without introducing strategic instabilities that decrease the overall likelihood of effective cooperative response to complex social problems.

While there is some reason to be confident in these predictions, given that they derive from well understood game theoretic models, they are currently unsupported by experimental observation. An important area for future research will be into the behaviour of "bio-enhanced" individuals in settings that contain opportunities for agents to adopt exploitative strategies in a variety of non-transparent games.

3. Complex interventions for moral bioenhancement

Thus far we have considered unilateral enhancement of overtly altruistic dispositions such as a tendency to empathise with others or to trust others. A more sophisticated strategy of moral bioenhancement may employ simultaneous adjustment of multiple dispositions. Perhaps by increasing our tendency to trust, while also enhancing our propensity for fairness (e.g. our retributive and reputation-monitoring tendencies), it will be possible to realise benefits of cooperation without additional exploitation.
Simultaneously modifying two or more psychological dispositions will likely have novel and unpredictable effects. On existing evidence, however, we have substantial reason to doubt that multi-dimensional moral bioenhancement will lead to straightforward benefits for solving global challenges like climate change (summary in Table 1).

Moral disposition	Negative effects of bioenhancement
Trust	Improves payoff for deceptive/exploitative strategies
Empathy	Improves payoff for exploitative strategies
Retribution	Lowers efficiency; may suppress cooperation
Reputation monitoring	Increased rate of false positives may suppress cooperation
Parochialism	Reduces out-group cooperation

Table 1. Summary of psychological dispositions relevant to cooperative behaviour and anticipated negative effects of enhancing those dispositions.

Consider the three defensive dispositions identified above: parochialism, reputation monitoring, and retribution.

Increasing *parochial tendencies* will no doubt assist in guarding against exploitation, but it will also drastically reduce the scope for *out-group* cooperation. In the context of global challenges that require cooperation between distinct groups, this therefore appears to be a hopeless suggestion.

Increasing *retributive tendencies* is likely to be of no benefit, or outright harmful. Models show that punishment of defectors can promote the evolution of cooperation under a variety of circumstances: if the interactions are structured by a network [50]; if interactions are non-anonymous [51]; if punishment is coordinated [52]; and if exit from the population is viable [53]. But many of these models also allow the evolution of so-called "anti-social punishment" – punitive behaviour directed towards cooperators [52-54] – casting doubt on the uniformly beneficial role of retributive behaviours in stabilising cooperation [55]. Some of these models also find that the effect of punishment is sensitive to the ratio of cost of punishment to the benefits of cooperation, but is relatively insensitive to the harshness of the punishment [51,52]. So increasing the harshness of the punishment to the defector is not robustly predicted to improve the levels of cooperation. Experiments also suggest that increasing the severity of punishment, even where it increases rates of cooperation, is likely to harm overall efficiency: groups with harsh punishment may cooperate more, but they are poorer [31,55]. Further, individuals who administer punishment tend to gain fewer benefits of cooperation than those who refrain [55]. Given that the aim of moral bioenhancement is to seek maximally group beneficial solutions, using harsher penalties for transgression would appear to be self-undermining.

Reputation monitoring is a more promising avenue for intervention. The success of online trading platforms such as eBay has been premised on the establishment of trust between relatively anonymous trading partners, enabled by mechanisms for the tracking of reputations [56]. But this example reinforces the point that the primary obstacles to cooperation are social and institutional, rather than cognitive. Furthermore, we can find no evidence that reputation monitoring could be made more effective by pharmacological or genetic interventions. It is conceivable that reputation-monitoring could be promoted by some sort of intervention which lowers the evidential threshold before judging another agent to have defected. (For instance, there is some evidence that modafinil induces overconfidence in visuomotor abilities in fatigued subjects [57]. Perhaps future drugs could be synthesised to induce overconfidence in the

domain of judging others to have transgressed in a cooperative setting.) But this will increase the rate at which false positives occur, and consequently agents will more frequently be punished "unfairly" for pro-social behaviour. Experimental evidence suggests that punishment of cooperators is highly destructive of cooperation [58]. Moreover, models of group-structured populations suggest that the destructive effects of punishing cooperators increase with the severity of the punishment [59].

4. Conclusion

Insofar as the stimulus for moral bioenhancement issues from concerns about human societies' ability to cooperatively tackle problems for which our evolutionary inheritance has not ideally equipped us, it needs to come furnished with some kind of theory about what makes effective cooperation achievable in the face of grave risks and collective action problems. Moral bioenhancement that simply purports to make us more trusting, generous, empathetic, etc., overlooks the ways in which cooperative success relies on a complex network of social dispositions – some of them involving the imposition of net social costs. And it fails to account for the danger that, by tinkering with some of these dispositions, we could have an adverse effect on other dispositions, or on the equilibria that obtain between different dispositions.

This conclusion does not entail pessimism. There may well be radical innovations that could help human societies engage in cooperative problem-solving more effectively than at present. But if we are to enhance human cooperation we should start with a good theory about what makes our existing cooperative endeavours possible in the first place.

Acknowledgements

Thanks to Chet Pager and Janine Perlman for helpful feedback on earlier drafts.

References

1 Overpeck JT, Meehl GA, Bony S, *et al.* Climate Data Challenges in the 21st Century. *Science* 2011;**331**:700–2. doi:10.1126/science.1197869

2 Loewenstein G, Elster J. *Choice over Time*. New York: : Russell Sage Foundation 1992.

3 Gifford R. The dragons of inaction: Psychological barriers that limit climate change mitigation and adaptation. *Am Psychol* 2011;**66**:290–302. doi:10.1037/a0023566

4 Markowitz EM, Shariff AF. Climate change and moral judgement. *Nature Climate Change* 2012;**2**:243–7. doi:doi:10.1038/nclimate1378

5 Broome J. The ethics of climate change. *Scientific American* 2008;**298**:96–102.

6 Gardiner SM. *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. Oxford: : Oxford University Press 2011.

7 Persson I, Savulescu J. *Unfit for the Future: The Need for Moral Enhancement*. Oxford: : Oxford University Press 2012.

8 Douglas T. Moral Enhancement. *Journal of Applied Philosophy* 2008;**25**:228–45. doi:10.1111/j.1468-5930.2008.00412.x

9 Persson I, Savulescu J. Getting moral enhancement right: the desirability of moral bioenhancement. *Bioethics* 2013;**27**:124–31.

10 Sparrow R. Better Living through Chemistry? A Reply to Savulescu and Persson on "Moral Enhancement." *Journal of Applied Philosophy* 2014;**31**:23–32.

11 Sparrow R. Egalitarianism and moral bioenhancement. *American Journal of Bioethics* 2014;**14**:20–8.

12 Harris J. Moral enhancement and freedom. *Bioethics* 2011;**25**:102–11.

13 Sorensen K. Moral enhancement and self-subversion objections. *Neuroethics* 2014;**7**:275–86.

14 Bublitz C. Moral enhancement and mental freedom. *Journal of Applied Philosophy* 2015;**32**:88–106.

15 Levy N, Douglas T, Kahane G, *et al.* Are you morally modified?: the moral effects of widely used pharmaceuticals. *Philosophy, psychiatry, and Psychology* 2014;**21**:111–25.

16 Douglas T, Bonte P, Focquaert F, *et al.* Coercion, incarceration, and chemical castration: An argument from autonomy. *Bioethical Inquiry* 2013;**10**:393–405.

17 Sally D. Conversation and Cooperation in Social Dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society* 1995;**7**:58–92.

18 Hamilton WD. The genetical evolution of social behaviour. II. *Journal of Theoretical Biology* 1964;**7**:17–52.

19 Trivers RL. The Evolution of Reciprocal Altruism. *Quarterly Review of Biology* 1971;**46**:35–57.

20 Nowak MA, Sigmund K. Evolution of indirect reciprocity. *Nature* 2005;**437**:1291–8. doi:10.1038/nature04131

21 Bowles S, Gintis H. *A Cooperative Species*. Princeton: : Princeton University Press 2011.

22 Bernhard H, Fischbacher U, Fehr E. Parochial altruism in humans. *Nature* 2006;**442**:912–5. doi:10.1038/nature04981

23 Choi JK, Bowles S. The Coevolution of Parochial Altruism and War. *Science* 2007;**318**:636–40. doi:10.1126/science.1144237

24 Fu F, Tarnita CE, Christakis NA, *et al.* Evolution of in-group favoritism. *Sci Rep* 2012;**2**:460. doi:10.1038/srep00460

25 Nowak MA, Sigmund K. Evolution of indirect reciprocity by image scoring. *Nature* 1998;**393**:573–7.

26 Barclay P. Reputational benefits for altruistic punishment. *Evol & Hum Beh* 2006;**27**:325–44.

27 Brandt H, Sigmund K. Indirect reciprocity, image scoring, and moral hazard. *Proc Natl Acad Sci USA*

2005;**102**:2666–70.

28 Shinada M, Yamagishi T. Punishing free riders: direct and indirect promotion of cooperation. *Evol & Hum Beh* 2007;**28**:330–9.

29 Gintis H. Punishment and Cooperation. *Science* 2008;**319**:1345–6. doi:10.1126/science.1155333

30 Fehr E, Fischbacher U. Third-party punishment and social norms. *Evol & Hum Beh* 2004;**25**:63–87.

31 Egas M, Riedl A. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences* 2008;**275**:871–8. doi:10.1126/science.1065507

32 Batson CD, Ahmad N, Lishner DA. Empathy and Altruism. In: Snyder CR, Lopez SJ, eds. *The Oxford Handbook of Positive Psychology*. Oxford: : Oxford University Press 2011. 417–26.

33 Hu Y, Strang S, Weber B. Helping or punishing strangers: neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Front Behav Neurosci* 2015;**9**:1–11. doi:10.3389/fnbeh.2015.00024

34 Thielmann I, Hilbig BE. The Traits One Can Trust: Dissecting Reciprocity and Kindness as Determinants of Trustworthy Behavior. *Personality and Social Psychology Bulletin* 2015;**41**:1523–36.

35 Glaeser EL, Laibson DI, Scheinkman JA, *et al.* Measuring trust. *The Quarterly Journal of Economics* 2000;**115**:811–46.

36 Dohmen T, Falk A, Huffman D, *et al.* The Intergenerational Transmission of Risk and Trust Attitudes. *The Review of Economic Studies* 2012;**79**:645–77. doi:10.1093/restud/rdr027

37 Johnson ND, Mislin AA. Trust games: A meta-analysis. *Journal of Economic Psychology* 2011;**32**:865–89.

38 Bicchieri C, Xiao E, Muldoon R. Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy & Economics* 2011;**10**:170–87.

39 Falk A, Zehnder C. A city-wide experiment on trust discrimination. *Journal of Public Economics* 2013;**100**:15–27.

40 Zak PJ, Stanton AA, Ahmadi S. Oxytocin increases generosity in humans. *PLoS ONE* 2007;**2**:e1128. doi:10.1371/journal.pone.0001128

41 Kosfeld M, Heinrichs M, Zak PJ, *et al.* Oxytocin increases trust in humans. *Nature* 2005;**435**:673–6. doi:10.1038/nature03701

42 Baumgartner T, Heinrichs M, Vonlanthen A, *et al.* Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 2008;**58**:639–50.

43 De Dreu CKW, Greer LL, Handgraaf MJ, *et al.* The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science* 2010;**328**:1408–11.

44 De Dreu CKW, Greer LL, Van Kleef GA, *et al.* Oxytocin promotes human ethnocentrism. *Proc Natl Acad Sci USA* 2011;**108**:1262–6.

45 Butler J, Giuliano P, Guiso L. The right amount of trust. NBER Working Papers. 2009.

46 Dietz T, Ostrom E, Stern PC. The struggle to govern the commons. *Science* 2003;**302**:1907–12.

47 Skyrms B. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: : Cambridge University Press 2004.

48 Baumard N, André J-B, Sperber D. A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences* 2013;**36**:59–78.

49 DeCanio SJ, Fremstad A. Game theory and climate diplomacy. *Ecological Economics* 2013;**85**:177–87. doi:10.1016/j.ecolecon.2011.04.016

50 Nakamaru M, Iwasa Y. The evolution of altruism by costly punishment in lattice-structured populations: score-dependent viability versus score-dependent fertility. *Evolutionary Ecology Research* 2005;**7**:853–70.

51 Hilbe C, Traulsen A. Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Sci Rep* 2012;**2**. doi:doi:10.1038/srep00458

52 Boyd R, Gintis H, Bowles S. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 2010;**328**:617–20.

53 Hauert C, Traulsen A, Brandt H, *et al.* Via Freedom to Coercion: The Emergence of Costly Punishment. *Science* 2007;**316**:1905–7. doi:10.1126/science.1141588

54 Nakamaru M, Sasaki A. Can transitive inference evolve in animals playing the hawk–dove game? *Journal of Theoretical Biology* 2003;**222**:461–70.

55 Dreber A, Rand DG, Fudenberg D, *et al.* Winners don't punish. *Nature* 2008;**452**:348–51.

56 Resnick P, Zeckhauser R. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. In: Baye MR, ed. *The Economics of the Internet and E-commerce*. Elsevier Science 2002. 127–57.

57 Repantis D, Schlattmann P, Laisney O. Modafinil and methylphenidate for neuroenhancement in healthy individuals: a systematic review. *Pharmacological Research* 2010;**62**:187–206.

58 Gächter S, Herrmann B. The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review* 2011;**55**:193–210.

59 Powers ST, Taylor DJ, Bryson JJ. Punishment can promote defection in group-structured populations. *Journal of Theoretical Biology* 2012;**311**:107–16. doi:10.1016/j.jtbi.2012.07.010