Modelling the social dynamics of moral enhancement: social strategies sold over-the-counter and the stability of society*

Joao Fabiano, M.Phil.

Faculty of Philosophy, University of Oxford

Oxford Uehiro Centre for Practical Ethics

Littlegate House, Suite 7, Room 6, St Ebbes Street, Oxford, OX1 1PT, United Kingdom

+44 (0)1865 286888.

joao.fabiano@philosophy.ox.ac.uk

Anders Sandberg, Ph.D.

James Martin Research Fellow

Oxford Martin School; Faculty of Philosophy, University of Oxford

Future of Humanity Institute

Littlegate House, Suite 1, St. Ebbes Street, Oxford, OX1 1PT, United Kingdom

+44 (0) 1865 286800

anders.sandberg@philosophy.ox.ac.uk

Abstract: How individuals tend to evaluate the combination of their own and other's payoffs – Social Value Orientations – is likely to be a potential target of future moral enhancers. However, the stability of cooperation in human societies has been buttressed by evolved mildly prosocial orientations. If they could be

* **Acknowledgements:** The ideas behind this paper were first discussed in a seminar and subsequent discussion at the Centre for Research in Social Simulation in Surrey, where Jen Badham, Juan Cano and Corinna Elsenbroich helped properly shaping our models and assumptions. The paper also benefited from discussions with Brian Earp on cooperation and individualism. Emma Bates and Paulo Salem gave much appreciated comments to improve clarity. We have also benefited from the comments and questions from the attendees of the Belgrade "Enhancing the understanding of enhancement" conference October 27-28 2015.

changed, would this destabilize the cooperative structure of society?

We simulate a model of moral enhancement where agents play games with each other and can enhance their orientations based on maximizing personal satisfaction. We find that given the assumption that very low payoffs lead agents to be removed from the population, there is a broadly stable prosocial attractor state. However, the balance between prosociality and individual payoff-maximization is affected by different factors. Agents maximizing their own satisfaction can produce emergent shifts in society that reduce everybody's satisfaction. Moral enhancement considerations should take the issues of social emergence into account.

Keywords: moral enhancement, social value orientation, prosociality, selfishness, altruism, game theory, computer simulation, emergence

1. Moral Enhancement: desirable, feasible and poorly forecasted

The decisions we make when we establish social relationships, from dating to marriage, and from small group organization to global coordination, often deal with conflicts between the individual and society, which we refer to as social dilemmas. They can be understood as situations in which it is tempting for each individual to take a non-cooperative course of action that is individually better in the short term, but that, should everyone take that course, would make everyone

worse off in the long term[1]. This area of research benefits heavily from game theory, but empirically demonstrates that individuals will choose according to preferences for specific combinations of outcomes; outcomes that are contrary to the classical *equilibria* produced by pure rational choices in game theory. Rather than solely aiming at maximizing their own payoff, individuals will choose particular combinations between their and others' payoffs[2]. For instance, often individuals prefer solutions where the sum of everyone's payoffs is maximized, instead of his/her own payoff (prosocial choices); or prefer solutions where the difference between each payoff is minimal (egalitarian choices); or, at times, solutions where others' payoffs are minimized, regardless of the potential harmful consequences for themselves (aggressive choices)[3]. With the empirical results of that area, we can model the *individual strategies* for social behaviour and then the dynamics of social interactions.

Over the course of the last century, humanity's inability to cooperate on an international scale has become a major concern, particularly when problems on the scale of global warming and nuclear disarmament have arisen. Ingmar Persson and Julian Savulescu have argued that we are not equipped with the right set of traits and morals to solve this problem[4]. They observe that our ability to cooperate well in extremely large groups, spread across countries and territories or from different ethnicities and backgrounds, is very limited. Additionally, we are gaining an ever-increasing destructive power and technology is rapidly becoming globalised, so that the probability of any particular individual having enough power to destroy the whole of humanity has increased. Therefore, the two authors conclude, we have a moral imperative to pursue moral enhancement, i.e.,

the improvement of our moral dispositions. Not doing so will expose humanity to extreme risks of catastrophes or extinction – what they call *ultimate harm.*

Instead of focusing on the moral obligation that society has to promote the development of moral enhancement, Tom Douglas has analysed the moral permissibility of a single individual voluntarily performing moral enhancement[5]. According to Douglas, while many forms of human enhancement are sometimes considered morally impermissible on grounds that they produce an advantage for the individual at the cost of a disadvantage for society, these same grounds could not be applied to moral enhancement. If a person enhances herself so that she will have morally better motives, or so that her actions conform better to common moral expectations, then this has clear advantages to society as a whole. Moral enhancement seems hard to oppose, probably desirable and perhaps pursuing it is a moral imperative.

Although the near future feasibility of such radical manipulation of human moral dispositions remains uncertain, recent studies have demonstrated some aspects of cooperation and moral judgment are subject to pharmacological manipulation. For instance, serotonin seems to be positively correlated with cooperation; serotonin depleted individuals are more likely to continue overharvesting a common resource pool[6]. Furthermore, exogenous oxytocin administration was demonstrated to be correlated with trust[7], generosity[8], empathy[9] and several other traits related to co-operation[10]. Common variants in the oxytocin receptor seem to underlie partially individual differences in prosociality[11]. Therefore, it seems reasonable to assume we might come to develop and use drugs which will change our social preferences in the future, thus dramatically influencing all our social

interactions. Moreover, as John Shook contends, this development might happen regardless of possible unresolved conceptual issues in the moral enhancement debate[12].

Having the power to choose freely previously unwilled innate fundamental social preferences will introduce new and powerful dynamics in society, for better or worse. However, not only do we still have merely a crude understanding of how those social preferences currently work in large societies, but we also lack a model to help predict what will happen once we start changing those preferences with the use of enhancement technologies. Our primary goal here will be to construct the first of such models and indicate future research directions in this area.

A secondary goal will be addressing a common worry about moral enhancement. Many critics argue that while the problems moral enhancement proposes to solve are mainly political and social, moral enhancement focuses solely on the individual. To increase an individual predisposition to act pro-socially or to empathize would do little to solve problems that are structural and at a societal level[13]. David Wasserman draws attention to the fact that a universally morally enhanced society might not be functional[14]. Masahiro Morioka and John Shook mention that the morally enhanced could be easier targets for domination due to decreased aggression and increased tendency to cooperate[15]. By modelling the dynamics of a whole population of agents that can freely choose their social preferences, we hope to reveal some possible population-level effects from the introduction of moral enhancement technologies targeting the behaviour of individuals.

2. Prerequisites

2.1 Basic Assumptions

We will start with a few simple uncontroversial background assumptions. They come from the science of diffusion of innovations (1 and 2), and experimental social psychology (3).

1. **Early-adopters:** Given the introduction of a potentially beneficial new technology, at least a small group of risk-taking early-adopters will make use of it[16].
2. **Imitation**: An agent will be more likely to adopt a technology that increases a certain social preference if she thinks that this social preference is successful, which she will assess by seeing how well, through her eyes, other agents with that social preference do[17].
3. **Social Value Orientations**: Social preferences or strategies correspond to *individual preferences* over certain specific combinations of payoffs to oneself and others[18]. Those preferences can be mapped by social value orientations (SVO), illustrated by the following representation:

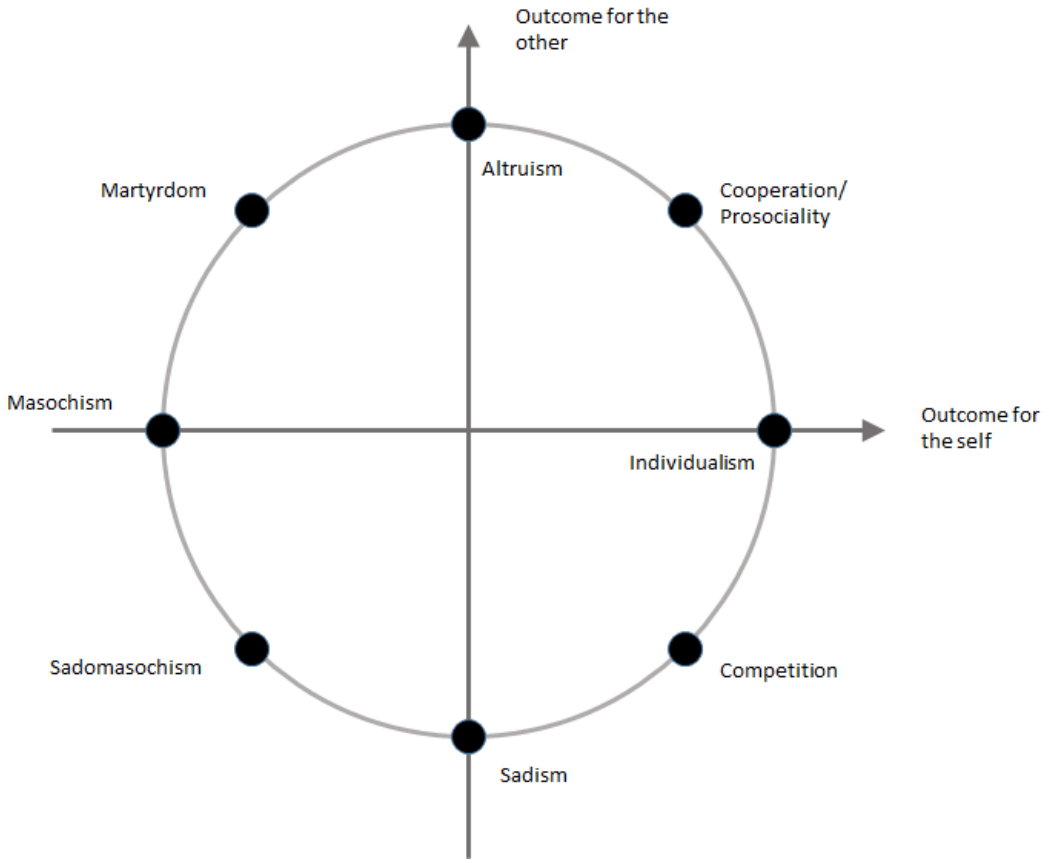

Figure 1: Social Value Orientations Ring Measure

Each agent will have an SVO, which positions her in the plotted ring with regards to her preferences for her own outcome and the outcomes of others. A cooperative agent would have equal preferences for positive outcomes for herself and others; a competitive agent would have a preference for positive outcomes for herself and negative for others; and an individualist agent would have a preference for positive outcomes for herself and no preference over others' outcome (the terms are taken from the SVO literature[19]).

In experimental measurements of human SVO, the majority lies between individualistic or prosocial, with some competitive individuals[20]. We will call this distribution the "typically human" below.

2.2 How do I choose when I am choosing how I choose?

Our subject matter requires one new background assumption. Unlike in diffusion of innovations, our agents do not have stable preferences; instead, they are choosing those preferences. Unlike the evolution of cooperation, the compelling selection force does not come from slow, millennial and stable evolutionary pressures; they come from agents' choices. Finally, although experimental social psychology often accounts for some flexibility, we are – *ex hypothesis* – assuming drugs that would vastly increase our freedom to choose willingly those orientations. Here we have an unusual degree of freedom over the decision processes themselves. We need to model how agents choose when they are choosing strategies; strategies that will, in turn, determine how they will choose in the future.

While it is possible that some intellectuals will rationally sit down and evaluate what preferences they ought to have (see section 4), in practice, the updating is likely to be piecemeal and affected by what other agents do.

We will model a society where randomly selected pairs of agents interact with each other, the interactions producing different outcomes defined by a 2x2 payoff matrix (which can either be random or correspond to a standard game such as the Prisoner's Dilemma). The agents receive *satisfaction scores* which are their utility evaluations of outcomes. An agent’s utility is the result of his own payoff, the other’s payoff, and the agent’s SVO – i.e. the agent’s preferences for each payoff. After a given amount of rounds, some agents will update their SVO based on the satisfaction score of other agents.

A simple way of updating is if an agent tries to imitate agents with higher scores (objective imitation). However, this would neglect the fact that the other agent's score is the result of an evaluation of how well it did with respect to its own utility function. Imitating agents with higher satisfaction scores will not necessarily yield a high score *according to the present utility function* of the imitator agent. A more elaborate way to update is if agents imitate agents who did best *according to the utility function of the imitator* (subjective imitation). Empirical data suggests that real-world situations will be likely to contain mixtures of both imitation strategies, with higher weight given to one's own standard of success[21].

3. Simulation

3.1. Model

Building on the theoretical assumptions of the last section we constructed a computer simulation using Matlab™ of the spread of social preference change among simple agents.

In the simulation each individual agent has a weight vector $\boldsymbol{w}$ of SVO values, corresponding to how strongly their preferences align with each orientation. The resulting weighting is used to evaluate the utility of different situations. Following Jeff Joireman et al., each agent has a *vitality score*, which simply measures her total individual payoff, and a *satisfaction score*, which measures the agent's overall utility according to his SVO[22]. For instance, a fully altruistic agent's vitality score would equal her own payoff while her satisfaction score would equal the other agent's payoff. Given a payoff for self (S) and other (O),

the satisfaction is calculated as $S^*(\boldsymbol{w}\cdot\boldsymbol{\alpha})+O^*(\boldsymbol{w}\cdot\boldsymbol{\beta})$ where $\boldsymbol{\alpha}$ is the vector of SVO weightings of self and $\boldsymbol{\beta}$ of other (e.g. $\alpha_{altruist}=0$, $\beta_{altruist}=1$; while $\alpha_{prosocial}=0.5$, $\beta_{prosocial}=0.5$)[23]. $\boldsymbol{w}$ is normalized, $|\boldsymbol{w}|=1$.

The standard simulation began with 100 agents with SVO weightings set either randomly or to the human population norm[24]. The agents then went through 5000 epochs, each consisting of each agent playing ten two-person games with randomly selected agents. The games were either randomly generated 2x2 games (payoffs were independently distributed normal random numbers) or a classical game. The agents involved calculated their own satisfaction with the different possible outcomes, producing subjective payoff matrices and then chose an action according to a mixed strategy Nash equilibrium calculated using the Lemke-Howson algorithm. Depending on the outcome their total satisfaction and vitality were updated based on the subjective and objective payoff.

At the end of an epoch, an agent will compare his score with others and update his SVO towards more successful agents. There are four ways for selecting other agents for comparison. By random selection of other agents, by having certain celebrity agents being more likely as comparisons (agent number n has probability n/N of being chosen), by selecting agents within a certain distance $|\boldsymbol{w}_{self}-\boldsymbol{w}_{other}|<d$, or by finding the agent with the highest perceived success. Perceived success is calculated using the other agent's satisfaction (objective measure) and the imitator agent satisfaction with the other's outcomes (subjective measure). Then, the agent will update his SVO weight vector a fraction towards this other agent: $\boldsymbol{w}_{new}\leftarrow (1-r)\boldsymbol{w}_{old} + r^*\boldsymbol{w}_{other}$. The degree to which an agent will copy the other at this point is set by the imitation rate r. Additionally, there is a

1% probability of an SVO weighting being randomized, producing mutations as individuals sometimes (deliberately or accidentally) choose larger variations.

Finally, we added the option of a *poverty threshold* such that if an agent has a vitality score below a certain value, generally set at the bottom 5%, he is removed and replaced by a typical human agent or a copy of a random agent – agents cannot survive on subjective satisfaction alone.

3.2 Results

If there is no poverty threshold, randomly initialized agents playing random games will converge to two possible attractor states: one where the weight vectors on average point in a prosocial direction (Fig.2) and one where the weight vectors point in the sadomasochistic direction (Fig.3). These attractor states represent *consistent motivations* for all agents. A population full of Sadomasochists or Prosocial agents will produce nearly ideal outcomes as agents can consistently minimize or maximize the joint outcome, respectively, thus maximizing everyone's satisfaction. A population of agents with Martyr or Competitive orientations cannot produce win-win outcomes. Hence, the agents tend to evolve towards either the Sadomasochist or Prosocial attractor state.

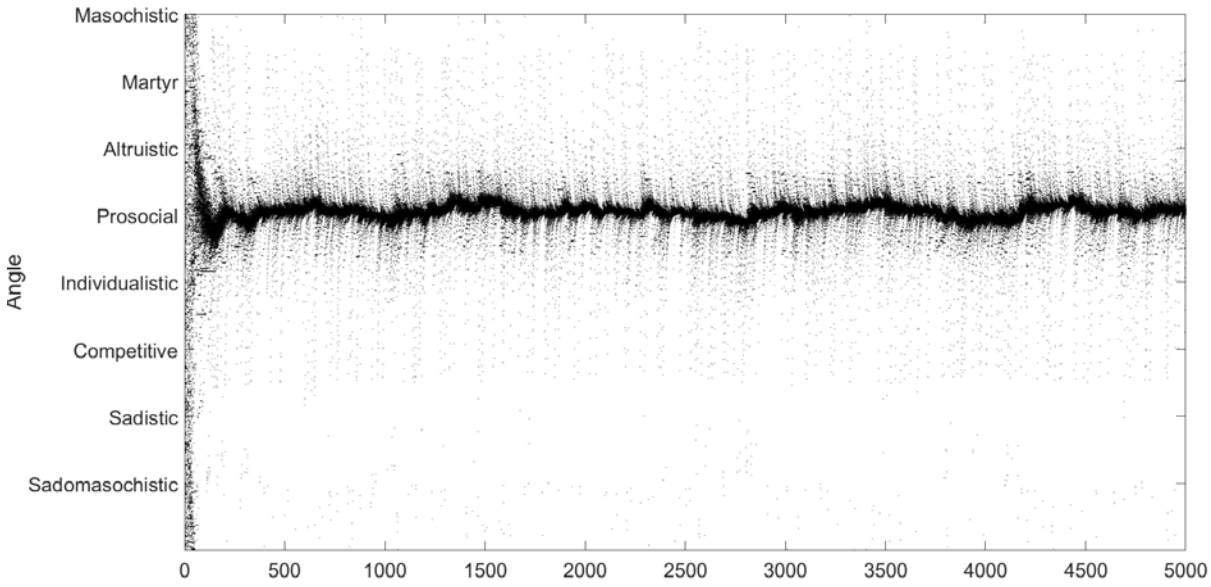

Figure 2: Prosocial attractor state - Each point localizes one agent's SVO across epochs

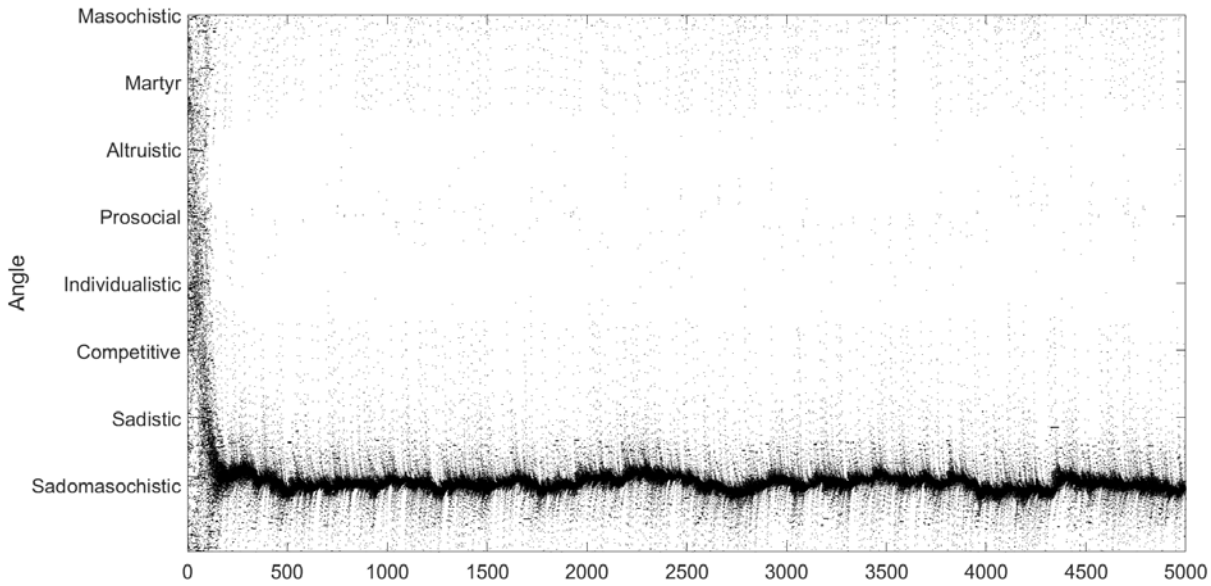

Figure 3: Sadomasochist attractor state

The prosocial attractor state is relatively close to the typically human initial state. It represents a state where agents aim at some mixture of positive outcomes for themselves and others. The negative attractor is incompatible with the long-term viability of the society given it produces very low objective payoffs. In the real

world, orientations that predispose towards it have probably been strongly selected against, evolutionarily and culturally. Next, we show how this baseline result is affected by changing the various parameters in our model.

Poverty threshold

We model the selection against agents with low vitality by introducing the poverty threshold with human replacement thereby replacing agents with vitality below a certain level with a typical human agent, which makes only the prosocial attractor stable[25]. Removing the agents in the bottom 5% of the vitality score distribution almost always destabilizes the negative attractor, while further increasing the threshold above 5% makes the prosocial attractor more and more robust. The greatest amount of variation is achieved with thresholds close to 1% in which the population is composed of a fluctuating mixture of prosocial and sadistic agents (Fig. 4), even when using the typical human initial state.

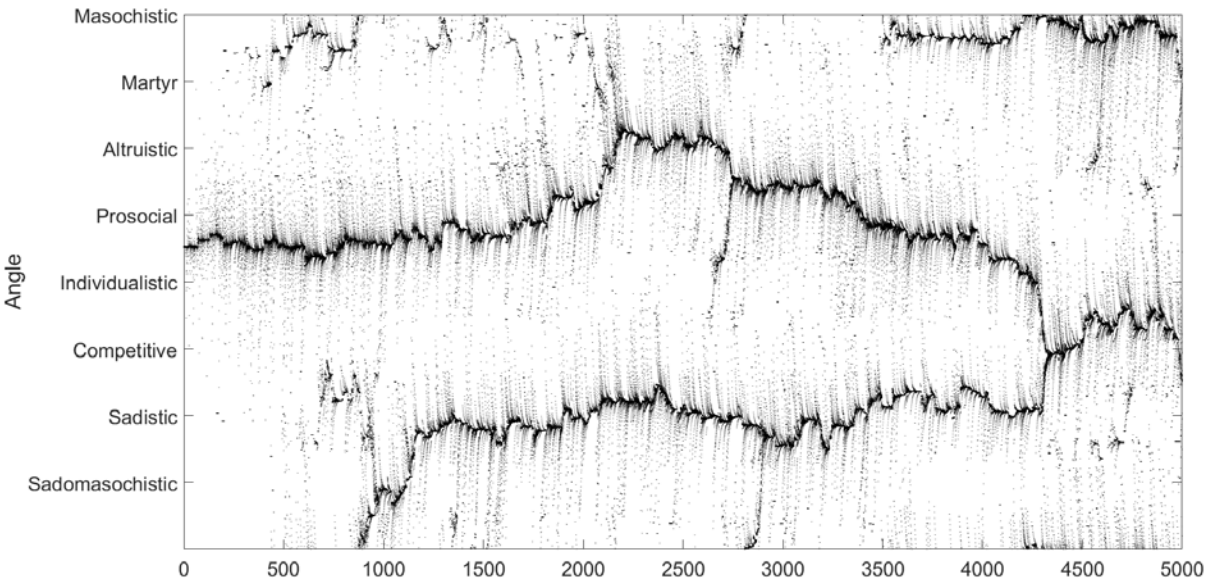

Figure 4: Mixed prosocial and sadomasochistic state

If instead of replacing agents below the threshold with a typical human agent we

use a random copy of another agent the results are very similar, but there is an increasing tendency for agents' orientations to become more varied while the population as a whole remains around the human average.

Available SVOs

It is not clear which and in which order actual social enhancers will be developed; thus, we have also modelled scenarios where agents cannot choose the full range of possible SVOs. If agents can only change either their individualistic or prosocial orientations, separately or in conjunction, then even with very small poverty thresholds the prosocial state is still fairly robust. When changing any other orientation becomes an option, agents increase in variability; nonetheless, the population as a whole remains prosocial for thresholds of 5% and above.

Game type

Besides the Prisoners' Dilemma (see below), we have also tested several other classical games such as the Chicken Game, Stag Hunt, Battle of Sexes, Pure Coordination, Odds and Evens and the Ultimatum Game. Except for the Chicken Game producing slightly more anti-social orientations, most of our results remained robust to the game being played.

Comparison mode

With regards to the comparison procedure, we have also tested random selection for comparisons, selecting nearby agents and selection based on agents' popularity. Whether comparison (and imitation) is based on interactions with randomly selected agents, nearby agents or celebrity agents also has little effect, although celebrities with outlier values can sometimes temporarily pull much of

the population in their direction.

A more surprising similarity was between objective imitation (imitating agents with high satisfaction regardless of their values), subjective imitation (imitating agents doing well as evaluated by the standards of the agent considering the change) and mixtures: there was no discernible difference in outcome. One possible explanation is that in most situations other agents will not have wildly different SVO from the observer, and hence, their own estimation of their success will be correlated enough with the observer's to be a good guide. While our framework allows radically different types of agents with incompatible evaluations (e.g. an Individualist trying to make sense of the actions of a Martyr) the imitative dynamics also lead to a convergence in how evaluations are done.

Other parameters

The level of rationality of agents − selecting actions by maximum average utility or calculating a Nash equilibrium mixed strategy − also did not significantly affect outcomes. The imitation rate affects diversity: high imitation rates produce populations dominated by a few orientations, shifting in a stepwise manner over time. However, it does not measurably change the dominant SVOs. Using either the human or the random initial state for agents' SVOs also did not affect outcomes.

Overall, the results above suggest our model is robust to changes in details of how the agents function, with the exception of the poverty threshold.

Prisoner's Dilemma

The Prisoner's Dilemma is perhaps the archetypal social dilemma, with a tension

between cooperative behaviour and the temptation to exploit fellow agents. When used in our model, there are transitions between states where most agents cooperate and states where most agents defect. Such shifts are commonly observed in simulations where strategies can evolve[26]. We observed a similar behaviour. In this case agent strategies are embodied in their SVO: populations dominated by Competitors or Individualists will tend to be uncooperative but can be invaded by Prosocial agents, while Prosocial and Altruist orientations push towards cooperation, but can be exploited by Individualists. An added complication is that agents that are being exploited may experience high satisfaction with the outcome if they have sufficiently negative regard for their own payoffs.

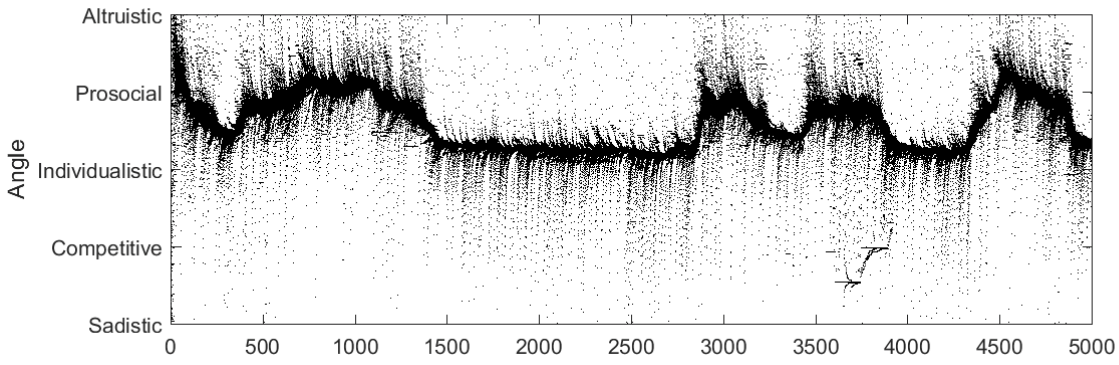

Figure 5: Dynamics of the Prisoner's Dilemma case.

Whether the emergent behaviour is cooperative depends on whether there are enough credible cooperators around, and once this threshold is crossed overall behaviour changes fast. In many cases, transitions were triggered by slow shifts in SVO continuing across the transition rather than any abrupt change in orientations. The level of average satisfaction in non-cooperative states can be

higher than in the start state, and sometimes satisfaction decreased during transitions to more cooperative states: agents adapted to a non-cooperative state were unhappy, despite gaining more objective payoffs.

This points to an important result of our model: just as individually maximizing payoffs can in some social situations produce suboptimal outcomes where everybody's outcome is worse, *individually maximizing satisfaction with outcomes can also lead to situations where everybody's satisfaction decreases.* The SVO framework decouples objective outcomes from experienced outcomes, but the problem of social dilemmas remains.

4. Complex imitation: choosing what to choose rationally

The above simulations have involved short-sighted individuals imitating each other or carrying out random changes. What if people actually rationally considered what orientation they ought to hold given the surrounding society? This section will consider a micro-model where a single agent considers its own weightings.

An agent can imagine himself with changed SVO weightings, and then evaluate the actions he would take based on these in different situations. The evaluation of whether such actions would be good would depend on the agent's current SVO[27]. For example, an altruistic agent can consider what he would do as a selfish individualist, judging the results based on the altruistic effects rather than what the potential future self would value. If the altruistic satisfaction in the scenario

would be higher than the current satisfaction, he might hence consider changing. Whether an agent would want to change depends both on the current values, the behaviour of other agents, and what kind of game-theoretic environment they find themselves in.

If society consists of typically human agents playing random games against each other, an individualist agent will on average do better if they have a more prosocial or altruist orientation. From the perspective of the agent seeking to maximize just their own satisfaction, it is hence rational to become more prosocial: there will be more individualist satisfaction in this situation because being motivated to enjoy other's outcomes will on average produce joint actions that have a higher direct payoff. This is also true for the individualist playing the Prisoner's Dilemma, since again more prosocial inclinations strongly increase their chances of mutually beneficial cooperation. Meanwhile, altruist agents find that they would do better if they include their own payoff in their utility calculation given this background. Prosocial agents are stable: they cannot do better if they update their orientation. This remains true even if the other agent orientations are random.

A far-sighted agent may consider whether they would update further. If the agent is an Individualist and playing random games against other random agents, it is rational to become Altruist. As an Altruist, it is rational to become an Individualist. The agent would hence flip between these states. If the agent begins as Prosocial in this scenario, it would not want to change. Prosocial hence seems to be uniquely stable, something that may explain the typically human orientation.

Note that such considerations may still lead to instability, especially for particular

social dilemma games. We have also just considered a single agent updating, assuming everybody else will remain the same. In general decisions about whether to change SVOs will happen under uncertainty about what other agents will do, both due to lack of knowledge of their states but also the computational complexity of predicting what other rational agents will do[28].

5. Discussion

In a population where agents can freely modify their social preferences and are not penalized for low payoffs, two scenarios are possible. Either the population will converge to a sadomasochist or a prosocial orientation. If, however, we introduce a payoff threshold below which agents are removed from the population, only the prosocial attractor remains stable. Small thresholds between zero and five percent of the payoff distribution produce mixed scenarios which seem to represent inviable societies. Limiting the available modification to only prosocial and individualistic orientations makes the prosocial state even more robust, even with very low thresholds. Making all modifications available tends to increase variation, but the overall population remains prosocial. Replacing agents randomly increases variation relative to replacing them with a typical human, but has a negligible effect on the overall population's orientation. Changing the game type, imitation style, rationality or imitation rate has little impact on the simulation.

It is hard to estimate what would be a realistic number for the poverty threshold; mortality does increase with decreasing income but in most developed countries rates still remain fairly low even at the very bottom of the distribution. However,

extreme poverty still functions as a strong demotivator in society. This seems to indicate the threshold should be below 5% and quite possibly below 1%, which would mean no single stable attractor once SVO modification technologies were introduced, except if we only introduce prosocial or individualistic modifications.

Which SVO modifications will be developed and used, and in which order, is hard to predict. The labelling, distribution and use of such technologies will depend on supply, demand and marketing. Shook envisions several catchy labels for technologies for enhancing thoughtfulness (Prudentia), moral beliefs (Ethicale), intentions (Benevolium), willpower (Prokrasia) and sensitivity (Sensitivia), which could be targeted at specific groups with different conceptions, and expectations, about morality[29]. These products would not even necessarily primarily modify the traits their names suggest. In our model, the spread is mainly determined by success, but prior perceptions and demand are likely to play a bigger role determining which modifications will be developed in the first place. Demand for moral enhancers is likely to be lower than for cognitive enhancers given that people associate moral traits as more fundamental and that empirical research has found that people are less likely to want to enhance fundamental traits such as kindness, empathy and self-control[30].

The replacement style will depend on the way these technologies will be deployed. Pharmacological interventions will probably mean a typically human replacement (e.g. an immigrant from another society or a person abandoning their pharmacological enhancements), whereas embryo selection or genetic engineering (perhaps targeting the oxytocin receptor gene) might mean random or other noisy replacements, which will create a more varied population.

Variation plays an important role in the dynamics: homogeneous populations can maintain a high and reliable degree of cooperation (e.g. in the Prisoner's Dilemma) but are vulnerable to invasion of outside non-cooperators because they lack subpopulations that resist the non-cooperators, while populations with a degree of internal variation also maintain such subgroups. If they are stuck in low cooperative states, they are conversely unable to escape them on their own. A low mutation rate or a high imitation rate produces fairly brittle societies: if social imitation in a moral enhancement scenario is too strong, it can be destabilizing. Moral monocultures may have the same ecological problems as biological monocultures.

It would seem that because of its self-consistency a broadly prosocial/individualistic orientation is likely to remain in society even when people can freely choose how to update their preferences. However, the balance between the prosocial and individualistic SVOs is sensitive to external incentives (such as the poverty threshold) and how agents copy each other. Even a small shift in average prosociality and individualism could have major social effects in human society, given the important effects existing differences in trust levels have on societies[31]. Since these differences are presumably due to non-biological factors at present[32], moral enhancement may induce far larger effects for good and ill. Moreover, if there is too low selection against low payoffs then the more likely scenario is a chaotic mixture of prosocial and sadomasochist orientations.

5.1. Future research directions

The agents modelled in this paper are very simple, and can be elaborated in various ways. We assumed agents were fully aware of each other’s preferences

and could solve game theoretic equilibria in their interactions: making them more boundedly rational can increase realism.

In real societies memory of past interactions and reputations have important effects on cooperation. A natural extension is to add memory and reputation to the model.

The societal structure can also be made more elaborate. For example, Wasserman points out that there might be benefits in having people with different orientation in different occupations (e.g. police) [33]. A natural exploration would be to identify some key functions in society that have different SVO requirements, the incentives for people in such positions, and see how stable the overall function would be. Another important issue to explore is how the enhancement influences different levels of group organisation and conflict.

5.2. Group-level effects

In many current economic and sociological theories, human society is a highly complex system whose organisation is partially (or primarily) determined by individual patterns of behaviour; patterns whose change can affect the system in unexpected ways. SVOs modelled here are likely to be one of the major individual patterns of behaviour shaping our overall society. We have limited our analysis to how enhancement technologies would alter those patterns at the individual level, but the effects these changes would have in group level interactions might actually go in the opposite direction.

In fact, groups that are highly cooperative internally will tend to be the least cooperative with other groups[34]. The relationship also holds in the opposite

direction: competition between groups leads to increased contribution to the public good within-group and to increased group effectiveness[35]. Men tend to exhibit higher levels of parochialism, cooperating more than women inside their group, but also have higher proclivity to conflict with out-groups[36].

Many theories have been proposed to explain why non-kin cooperation evolved, and several of them establish that this type of cooperation could only become evolutionarily stable if it had coevolved with aggression towards out-groups. For example, Bowles & Gintis attempted to model the evolution of cooperation using our best estimates regarding group-size and food-sharing during the Palaeolithic[37]. Even when using the most unfavourable estimates to this conclusion, their results show that parochialism and cooperation could only have evolved together[38].

Perhaps we might be able to decouple cooperation from inter-group competition by increasing prosociality and decreasing the sadistic orientation. Samuel Bowles and Herbert Gintis' model would suggest such populations would be extremely vulnerable: if they ever come into contact with parochial cooperators, they will lose every time. Other models such as Robert Boyd's altruistic punishment suggest the population would also easily fall prey to cheaters from the inside the population[39]. These group-level effects were not accounted for in our model, but are the next logical step for research in this area.

5.3. Conclusion

Our model suggests that shopping for social strategies as if freely choosing from a supermarket shelf does not automatically lead to utopia nor disaster. A generally

prosocial population is fairly stable, but we must consider three potential sources for risks. The level of variation inside the population can be affected by various factors − e.g., availability of modifications and poverty threshold − and it is not clear how this would affect overall society. The level of selection against low individual payoffs can affect the stability of a prosocial population, and there is a great deal of uncertainty about the real-world strength of this selection. Finally, there are many other factors not included in our current model, such as inter-group conflict and vulnerability to cheaters, that could lead to paradoxical effects even if individual agents become more prosocial.

Natural selection has set a group of constraints on what humans can be, and thus made defining ourselves and what we value easy by reducing our choices. We may use our inner aspirations and morals to define how things ought to be. However, once we break free from its chains we are overwhelmed by possibilities; we no longer have a set of common traits and preferences that must necessarily be held and whereby all other choices can be evaluated. Should our understanding of these issues be outpaced by our power of modifying ourselves, we risk extinction. As Nick Bostrom has stated, our accelerated technological development creates the necessity of philosophical inquiry with a deadline[40].

---

[1] Van Lange, P. A., Joireman, J., Parks, C. D., & Van Dijk, E. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*. 2013;*120*(2), 125-141.

[2] E.g. Van Lange, P. A. The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of personality and social psychology*. 1999;*77*(2), 337.

[3] Murphy, R. O., Ackermann, K. A., & Handgraaf, M. Measuring social value orientation. *Judgment and Decision Making*. 2011;*6*(8), 771-781.

[4] Persson, I., & Savulescu, J. The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy*. 2008;*25*(3), 162-177.
Persson, I., & Savulescu, J. *Unfit for the future: the need for moral enhancement*. Oxford University Press. 2012.

[5] Douglas, T. Moral enhancement. *Journal of applied philosophy*. 2008;25 (3), 228-245

[6] Bilderbeck, A. C., Brown, G. D., Read, J., Woolrich, M., Cowen, P. J., Behrens, T. E., & Rogers, R. D. Serotonin

and Social Norms Tryptophan Depletion Impairs Social Comparison and Leads to Resource Depletion in a Multiplayer Harvesting Game. *Psychological science*. 2014;*25*(7), 1303-1313.

[7] Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. Oxytocin increases trust in humans. *Nature*. 2005;*435*(7042), 673-676.

[8] Zak, P. J., Stanton, A. A., & Ahmadi, S. Oxytocin increases generosity in humans. *PLoS one*. 2007;*2*(11), e1128.

[9] Shamay-Tsoory, S. G., Fischer, M., Dvash, J., Harari, H., Perach-Bloom, N., & Levkovitz, Y. Intranasal administration of oxytocin increases envy and schadenfreude (gloating). *Biological psychiatry*. 2009;*66*(9), 864-870.

[10] Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., et al. Oxytocin modulates neural circuitry for social cognition and fear in humans. *The Journal of neuroscience*. 2005;*25*(49), 11489-11493.

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological psychiatry*. 2009;*65*(9), 728-731.

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. Oxytocin improves “mind-reading” in humans. *Biological psychiatry*. 2007;*61*(6), 731-733.

Krueger, F., Parasuraman, R., Moody, L., Twieg, P., de Visser, E., McCabe, K., et al. Oxytocin selectively increases perceptions of harm for victims but not the desire to punish offenders of criminal offenses. *Social cognitive and affective neuroscience*. 2013;*8*(5), 494-498.

Guastella, A. J., Mitchell, P. B., & Mathews, F. Oxytocin enhances the encoding of positive social memories in humans. *Biological psychiatry*. 2008;*64*(3), 256-258.

Unkelbach, C., Guastella, A. J., & Forgas, J. P. Oxytocin selectively facilitates recognition of positive sex and relationship words. *Psychological Science*. 2008;*19*(11), 1092-1094.

Fischer-Shofty, M., Levkovitz, Y., & Shamay-Tsoory, S. G. Oxytocin facilitates accurate perception of competition in men and kinship in women. *Social cognitive and affective neuroscience*. 2012;nsr100.

[11] Israel, S., Lerer, E., Shalev, I., Uzefovsky, F., Riebold, M., et al. The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task. *PloS one*. 2009;*4*(5), e5535.

[12] Shook, J. R. Neuroethics and the Possible Types of Moral Enhancement. *AJOB Neuroscience*. 2012;*3*(4), 3–14. doi:10.1080/21507740.2012.712602

[13]de Melo-Martin, I., & Salles, A. Moral Bioenhancement: Much Ado About Nothing? *Bioethics.* 2014;9702, 124–131.

Murphy, T. F. Preventing Ultimate Harm as the Justification for Biomoral Modification. *Bioethics*. 2014;9702.

Harris, J. Moral progress and moral enhancement. *Bioethics*. 2013;27(5), 285–90.

[14] Wasserman, D. When bad people do good things: will moral enhancement make the world a better place? *Journal of Medical Ethics.* 2014;*40*(6), 374–5.

[15] Morioka, M. Some Remarks on Moral Bioenhancement. In *The Future of Bioethics: International Dialogues*. 2014.

[16] Rogers, E. M. *Diffusion of innovations*. Simon and Schuster. 2010.

[17] See note 16, Rogers 2010.

[18] See note 1, van Lange, et al. 2013

[19] Griesinger, D. W., & Livingston, J. W. Toward a model of interpersonal motivation in experimental games. *Behavioral science*. 1973.

[20]See note 3, Murphy, et al. 2011.

[21] Poole, M. E., Langan-Fox, J., & Omodei, M. Contrasting subjective and objective criteria as determinants of perceived career success: A longitudinal study. *Journal of Occupational and Organizational Psychology*. 1993;66(1), 39-54.

Abele, A. E., & Spurk, D. How do objective and subjective career success interrelate over time?. *Journal of Occupational and Organizational Psychology*. 2009;82(4), 803-824.

[22] Joireman, J. A., Shelley, G. P., Teta, P. D., Wilding, J., & Kuhlman, D. M. Computer simulation of social value

orientation: Vitality, satisfaction, and emergent game structures. In *Frontiers in social dilemmas research.* 1996; 289-310. Springer Berlin Heidelberg.

[23] In some simulations weights for MaxDiff, |S-O|, and MinDiff, -|S+O|, were included to model (anti)egalitarian orientations.

[24] Extracted from: See note 3, Murphy, et al. 2011.

[25] Using the Prisoner's Dilemma game instead of random games also destabilizes the sadomasochist attractor. Agents trying to minimize the joint payoff in this case have a coordination problem that destabilizes the attractor.

[26] Axelrod, R., & Hamilton, W. D. The evolution of cooperation. *Science*. 1981;211(4489), 1390-1396.

[27] Technically: let the agent have expected utility function *U(a,o)* that gives its satisfaction given its own action *a* and other's actions o (given the probabilities in its current environment). The agent will perform the optimal action *a** that maximizes *U(a,o)*. If the agent changes its utility function to *U'(a,o)* it will now take optimal actions *a*'*. These can be evaluated according to the *current* utility function *U*, and if *U(a*',o)-U(a*,o)>0* it is rational for the agent to change its utilities to match *U'*.

[28] In general, it is a NP complete (i.e. hard) to find the Nash equilibria with the highest social welfare in games between rational agents (Conitzer, V., & Sandholm, T. New complexity results about Nash equilibria. *Games and Economic Behavior*. 2008;*63*(2), 621-641.)

[29] See note 12, Shook 2012.

[30] Riis, J., Simmons, J. P., & Goodwin, G. P. Preferences for enhancement pharmaceuticals: The reluctance to enhance fundamental traits. *Journal of Consumer Research*. 2008;*35*(3), 495-508.

[31] Fukuyama F *Trust: the Social Virtues and Creation of Prosperity.* London: Free Press. 1996

Knack, S., & Keefer, P. Does social capital have an economic payoff? A cross-country investigation. *The Quarterly journal of economics*. 1997;1251-1288.

Wike, R., & Holzwart, K. Where trust is high, crime and corruption are low. *Pew Research Center*. 2008

[32] Van Lange, P. A. Generalized Trust Four Lessons From Genetics and Culture. *Current Directions in Psychological Science*. 2015;*24*(1), 71-76.

[33] See note 14, Wasserman 2014.

[34]Bornstein, G. Intergroup conflict: individual, group, and collective interests. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology.* 2003;Inc, 7(2), 129–145.

De Dreu, C. K. W. Social conflict. *Current Sociology*. 2013;61, 696–713.

[35] E.g. Cardenas, J. C., & Mantilla, C. Between-group competition, intra-group cooperation and relative performance. *Frontiers in Behavioral Neuroscience.* 2015;9(February), 1–9.

Puurtinen, M., & Mappes, T. Between-group competition and human cooperation. *Proceedings. Biological Sciences, The Royal Society.* 2009; 276(September 2008), 355–360.

Burton-Chellew, M. N., Ross-Gillespie, A., & West, S. a. Cooperation in humans: competition between groups and proximate emotions. *Evolution and Human Behavior.* 2010;31(2), 104–108.

Bornstein, G., Winter, E., & Goren, H. Experimental study of repeated team-games. *European Journal of Political Economy.* 1996;12(4), 629–639.

[36] Sidanius, J., & Veniegas, R. C. Gender and race discrimination: The interactive nature of disadvantage. Reducing Prejudice and Discrimination. *The Claremont Symposium on Applied Social Psychology*. 2000;47–69.

[37] Bowles, S., & Gintis, H. *A cooperative species: Human reciprocity and its evolution*. Princeton University Press. 2011.

[38] Their model revealed that: (1) groups with non-parochial co-operators have a disadvantage over other groups and thus would not have evolved in the first place, however; (2) groups with parochial co-operators, that are willing to sacrifice themselves fighting against out-groups in order to benefit their peers, have an evolutionary advantage and; finally, (3) merely parochial groups have a general disadvantage.

[39] Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*. 2003;*100*(6), 3531-3535.

[40] Bostrom, N. *Superintelligence: Paths, dangers, strategies*. Oxford University Press. 2014.