*Modeling the Social Dynamics of Moral Enhancement**

Social Strategies Sold Over the Counter and the Stability of Society

ANDERS SANDBERG and JOAO FABIANO

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 changed, would this destabilize the cooperative structure of society? We simulate a model of moral enhancement in which 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

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 noncooperative 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 benefit, individuals will choose particular combinations between their and others' benefits.[2] For example, often individuals prefer solutions where the sum of everyone's benefits is maximized, instead of their own benefit (prosocial choices);

The ideas behind this article 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 in properly shaping our models and assumptions. The article 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.

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 669751).

*In the original version of this article, both online and in print, the funding information was omitted. It has been added and an addendum has been published.

doi:10.1017/S0963180116001109

or prefer solutions where the difference between each benefit is minimal (egalitarian choices); or, at times, solutions where others' benefits are minimized, regardless of the potential harmful consequences for themselves (aggressive choices).[3] With the empirical results of that area, it is possible to model the *individual strategies* for social behavior, 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 globalized, so that the probability of any particular individual having enough power to destroy the whole of humanity has increased. Therefore, these two authors conclude, we have a moral imperative to pursue moral enhancement; that is, 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 analyzed the moral permissibility of a single individual voluntarily performing moral enhancement.[5] According to Douglas, whereas many forms of human enhancement are sometimes considered morally impermissible on the 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 people enhance themselves so that they will have morally better motives, or so that their actions conform better to common moral expectations, then this has clear advantages to society as a whole. Moral enhancement seems hard to oppose and is 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 that some aspects of cooperation and moral judgment are subject to pharmacological manipulation. For example, 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 cooperation.[10] Common variants in the oxytocin receptor seem to underlie partially individual differences in prosociality.[11] Therefore, it seems reasonable to assume that we might come to develop and use drugs that 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. The primary goal of this article will be to construct the first of such models and indicate future research directions in this area.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

A secondary goal will be addressing a common worry about moral enhancement. Many critics argue that whereas 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 prosocially 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 resulting from decreased aggression and increased tendency to cooperate.[15] By modeling 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 behavior of individuals.

Prerequisites

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: Agents will be more likely to adopt a technology that increases a certain social preference if they think that this social preference is successful, which they will assess by seeing how well, through their eyes, other agents with that social preference do.[17]
3) Social value orientations (SVO): Social preferences or strategies correspond to *individual preferences* over certain specific combinations of benefits to oneself and others.[18] Those preferences can be mapped by SVO, illustrated in Figure 1.

Each agent will have an SVO, which positions that agent in the plotted ring with regard to his or her preferences for his or her own outcome and the outcomes of others. Cooperative agents would have equal preferences for positive outcomes for themselves and others; competitive agents would have a preference for positive outcomes for themselves and negative for others; and individualist agents would have a preference for positive outcomes for themselves 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.".

How Do I Choose When I Am Choosing How I Choose?

This subject matter requires one new background assumption. Unlike in diffusion of innovations, the agents here do not have stable preferences; instead, they are choosing those preferences. Unlike the evolution of cooperation, the compelling

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

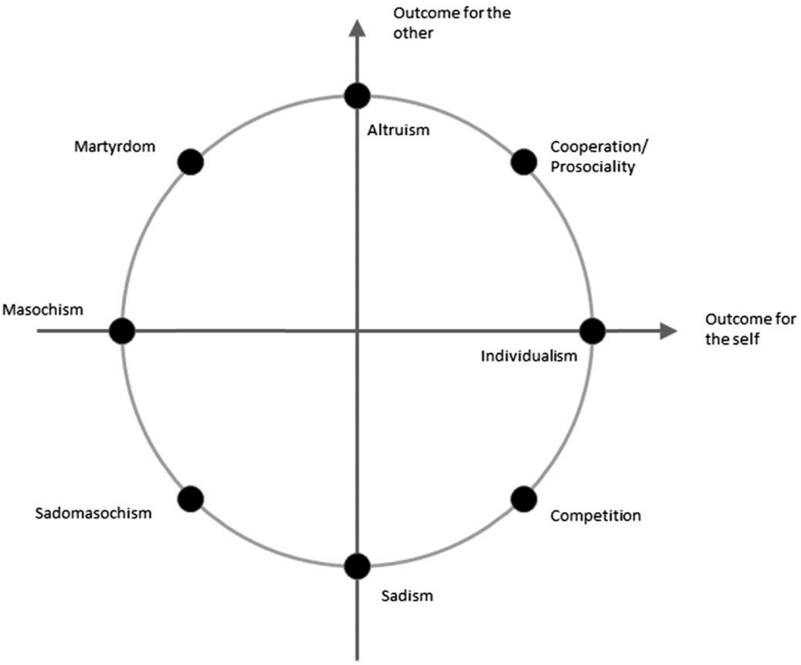

Figure 1. Social value orientations ring measure.

selection force does not come from slow, millennial, and stable evolutionary pressures; it comes 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 willingly choose those orientations. Here there is an unusual degree of freedom over the decision processes themselves. One will need to model how agents choose when they are choosing strategies; strategies that will, in turn, determine how they will choose in the future.

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

We will model a society in which randomly selected pairs of agents interact with each other, the interactions producing different outcomes defined by a 2 × 2 benefit matrix (which can either be random or correspond to a standard game such as the Prisoner's Dilemma). The agents receive *satisfaction scores* that are their utility evaluations of outcomes. An agent's utility is the result of his own benefit, the other's benefit, and the agent's SVO; that is, the agent's preferences for each benefit. 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 occurs when 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 occurs when agents imitate agents who did best *according to the utility function of the imitator* (subjective imitation).

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

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]

Simulation

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 that person's 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 the agent's total individual benefit, and a *satisfaction score,* which measures the agent's overall utility according to the agent's SVO.[22] For example, a fully altruistic agent's vitality score would equal that agent's own benefit, whereas the agent's satisfaction score would equal the other agent's benefit. Given a benefit 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 $\boldsymbol{\beta}$ is the vector of SVO weightings of other (e.g., $\alpha_{\text{altruist}} = 0$, $\beta_{\text{altruist}} = 1$; whereas $\alpha_{\text{prosocial}} = 0.5$, $\beta_{\text{prosocial}} = 0.5$).[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 10 two-person games with randomly selected agents. The games were either randomly generated 2 × 2 games (benefits were independently distributed normal random numbers) or a classical game. The agents involved calculated their own satisfaction with the different possible outcomes, producing subjective benefit 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 benefit.

At the end of an epoch, an agent would compare his or her score with others and update his or her SVO toward 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 $|\boldsymbol{w}_{\text{self}}\text{-}\boldsymbol{w}_{\text{other}}|<d$, 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 would update his or her SVO weight vector a fraction toward this other agent: $\boldsymbol{w}_{\text{new}}\leftarrow (1\text{-r})\boldsymbol{w}_{\text{old}} + \text{r}^*\boldsymbol{w}_{\text{other}}$. 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 percent 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 percent, that agent is

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

removed and replaced by a typical human agent or a copy of a random agent. Agents cannot survive on subjective satisfaction alone.

Results

If there was no poverty threshold, randomly initialized agents playing random games would converge to two possible attractor states: one in which the weight vectors on average pointed in a prosocial direction (Figure 2) and one in which the weight vectors pointed in the sadomasochistic direction (Figure 3). These attractor states represent *consistent motivations* for all agents. A population full of sadomasochists or prosocial agents would 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 toward either the sadomasochist or prosocial attractor state.

The prosocial attractor state is relatively close to the typically human initial state. It represents a state in which 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 that it produces very low objective benefits. In the real world, orientations that predispose toward 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 percent of the vitality score distribution almost always destabilizes the negative attractor, while further increasing

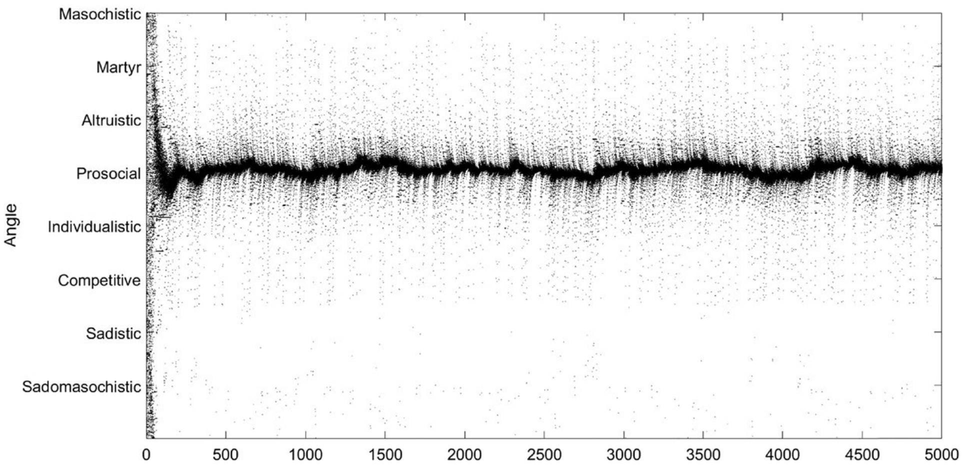

Figure 2. Prosocial attractor state: Each point localizes one agent's social value orientation (SVO) across epochs.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

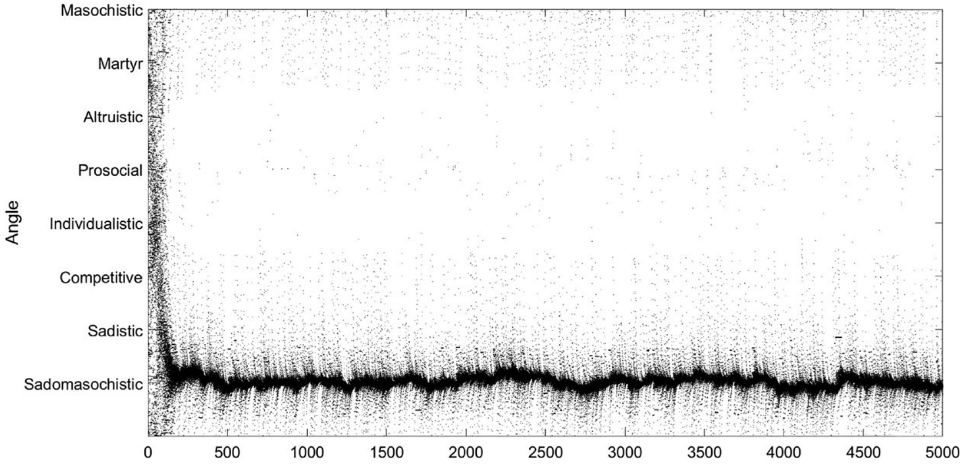

Figure 3. Sadomasochist attractor state.

the threshold above 5 percent makes the prosocial attractor more and more robust. The greatest amount of variation is achieved with thresholds close to 1 percent in which the population is composed of a fluctuating mixture of prosocial and sadistic agents (Figure 4), even when using the typical human initial 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; however, there is an increasing tendency for agents' orientations to become more varied while the population as a whole remains close to the human average.

Available SVOs

It is not clear which and in which order actual social enhancers will be developed; therefore, we have also modeled scenarios in which agents cannot choose the full

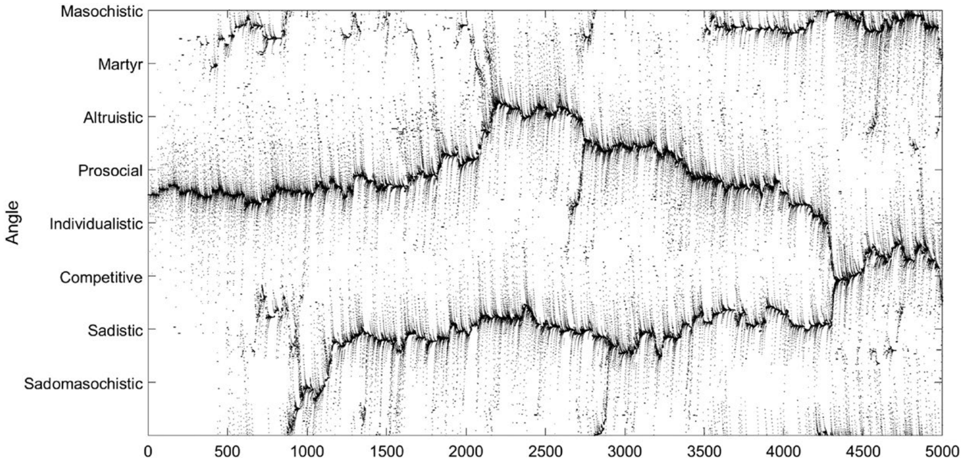

Figure 4. Mixed prosocial and sadomasochistic state.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

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 will still be 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 percent and above.

Game Type

In addition to Prisoners' Dilemma (see subsequent discussion), 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 antisocial orientations, most of our results remained robust to changes in the game being played.

Comparison Mode

With regard 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 a 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. Although 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 described suggest that our model is robust in the face of 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 behavior and the temptation to exploit fellow agents.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

When used in our model, there are transitions between states where most agents cooperate and states where most agents defect (Figure 5). Such shifts are commonly observed in simulations in which strategies can evolve.[26] We observed a similar behavior. 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, whereas prosocial and altruist orientations push toward cooperation, but can be exploited by individualists. An added complication is that agents who are being exploited may experience high satisfaction with the outcome if they have sufficiently negative regard for their own benefit.

Whether the emergent behaviour is cooperative depends on whether there are enough credible cooperators around, and once this threshold is crossed, overall behavior changes fast. In many cases, transitions were triggered by slow shifts in SVO continuing across the transition rather than by any abrupt change in orientations. The level of average satisfaction in noncooperative states can be higher than in the starting state, and sometimes satisfaction decreased during transitions to more cooperative states; agents adapted to a noncooperative state were unhappy, despite gaining more objective benefits.

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

Complex Imitation: Choosing What to Choose Rationally

The described simulations have involved short-sighted individuals, i.e., individuals imitating each other or making random changes without looking ahead at how they would fare in future interactions with other agents. What if people actually rationally considered what orientation they ought to hold given their expected interactions with surrounding society? This section will consider a micromodel in which a single agent considers his or her own weightings.

Agents can imagine themselves with changed SVO weightings, and then evaluate the actions they 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, altruistic agents can consider what they would do

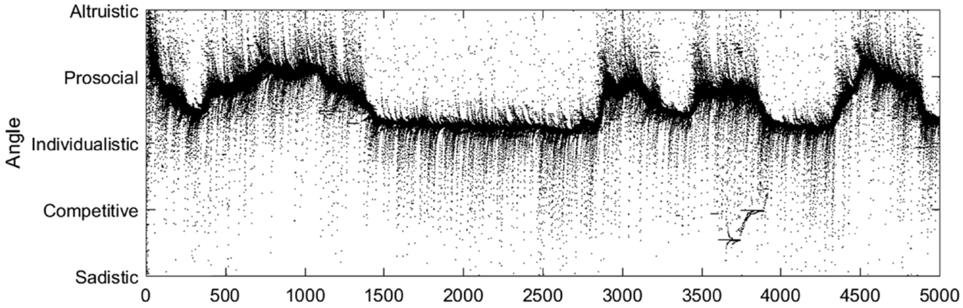

Figure 5. Dynamics of the Prisoner's Dilemma case.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

as selfish individualists, judging the results based on the altruistic effects rather than on what the potential future self would value. If the altruistic satisfaction in the scenario would be higher than the current satisfaction, the agent might hence consider changing. Whether certain agents would want to change depends on the current values, the behavior 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, individualist agents will on average do better if they have a more prosocial or altruist orientation. From the perspective of agents seeking to maximize only 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 others' outcomes will on average produce joint actions that have a higher direct benefit. This is also true for individualists playing the Prisoner's Dilemma, because again more prosocial inclinations strongly increase their chances of mutually beneficial cooperation. Meanwhile, altruist agents find that they would do better if they included their own benefit 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 agent looking at future interactions 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, that agent would not want to change. Prosocial hence seems to be a uniquely stable state, something that may explain the typically human orientation.

It is important to note that such considerations may still lead to instability, especially for particular social dilemma games. We have also only considered a single agent updating, assuming that everybody else will remain the same. In general, decisions about whether to change SVOs will happen when there is uncertainty about what other agents will do, not only because of lack of knowledge of their states, but also because of the computational complexity of predicting what other rational agents will do.[28]

Discussion

In a population in which agents can freely modify their social preferences and are not penalized for low benefits, two scenarios are possible: either the population will converge to a sadomasochist or a prosocial orientation. If, however, we introduce a benefit threshold below which agents are removed from the population, only the prosocial attractor remains stable. Small thresholds between 0 and 5 percent of the benefit distribution produce mixed scenarios that 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.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

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 that the threshold should be below 5 percent and quite possibly below 1 percent, which would mean no single stable attractor once SVO modification technologies were introduced, except if only prosocial or individualistic modifications were introduced.

It is hard to predict which SVO modifications will be developed and used, and in which order. 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 that their names suggest. In our model, the spread is mainly determined by success; however, prior perceptions and demand are likely to play a bigger role in 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 consider moral traits to be 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 pharmacological enhancements), whereas embryo selection or genetic engineering (perhaps targeting the oxytocin receptor gene) might mean random or other complex 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, whereas 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 that existing differences in trust levels have on societies.[31] Because these differences are presumably the result of nonbiological factors at present,[32] moral enhancement may induce far larger effects for good and ill. Moreover, if there is too low a selection against low benefits, then the more likely scenario is a chaotic mixture of prosocial and sadomasochist orientations.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

Future research directions

The agents modeled in this article are very simple, and can be elaborated in various ways. We assumed that 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 orientations in different occupations (e.g., police).[33] A natural exploration would be to identify some key functions in society that have different SVO requirements and the incentives for people in such positions, and to see how stable the overall function would be. Another important issue to explore is how the enhancement influences different levels of group organization and conflict.

Group-Level Effects

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

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 they also have a higher proclivity toward 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 have become evolutionarily stable if it had coevolved with aggression toward out-groups. For example, Bowles and Gintis attempted to model the evolution of cooperation using the best estimates regarding group size and food sharing during the Palaeolithic.[37] Even when using the most unfavorable estimates to this conclusion, their results show that parochialism and cooperation could only have evolved together.[38]

Perhaps it might be possible to decouple cooperation from intergroup competition by increasing prosociality and decreasing the sadistic orientation. Samuel Bowles and Herbert Gintis's model would suggest that 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 that the population would also easily fall prey to cheaters from 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.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

Conclusion

Our model suggests that shopping for social strategies as if freely choosing from a supermarket shelf does not automatically lead to either utopia or disaster. A generally prosocial population is fairly stable; however, there are three potential sources for risks. The level of variation inside the population can be affected by various factors—for example, availability of modifications and poverty threshold—and it is not clear how this would affect overall society. The level of selection against low individual benefits can affect the stability of a prosocial population, and there is much uncertainty about the real-world strength of this selection. Finally, there are many other factors not included in our current model, such as intergroup conflict and vulnerability to cheaters, which 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, therefore, 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 natural selection's 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]

Notes

1. Van Lange PA, Joireman J, Parks CD, Van Dijk E. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes* 2013;120(2):125–41.
2. For example, Van Lange PA. 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 RO, Ackermann KA, Handgraaf M. Measuring social value orientation. *Judgment and Decision Making* 2011;6(8):771–81.
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):162–77; Persson I, Savulescu J. *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press; 2012.
5. Douglas T. Moral enhancement. *Journal of Applied Philosophy* 2008;25(3):228–45.
6. Bilderbeck AC, Brown GD, Read J, Woolrich M, Cowen PJ, Behrens TE, et al. Serotonin and social norms tryptophan depletion impairs social comparison and leads to resource depletion in a multi-player harvesting game. *Psychological Science* 2014;25(7):1303–13.
7. Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E. Oxytocin increases trust in humans. *Nature* 2005;435(7042):673–6.
8. Zak PJ, Stanton AA, Ahmadi S. Oxytocin increases generosity in humans. *PLoS One* 2007;2(11):e1128.
9. Shamay–Tsoory SG, 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):864–70.
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):11,489–93; 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):728–31; Domes G, Heinrichs M, Michel A, Berger C, Herpertz SC. Oxytocin improves "mind-reading" in humans. *Biological Psychiatry* 2007;61(6):731–3; 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):494–8; Guastella AJ, Mitchell PB, Mathews F. Oxytocin enhances the encoding of positive social memories in humans. *Biological Psychiatry* 2008;64(3):256–8; Unkelbach C,

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press

Guastella AJ, Forgas JP. Oxytocin selectively facilitates recognition of positive sex and relationship words. *Psychological Science* 2008;19(11):1092–4; Fischer–Shofty M, Levkovitz Y, Shamay–Tsoory SG. Oxytocin facilitates accurate perception of competition in men and kinship in women. *Social Cognitive and Affective Neuroscience* 2013;8:313–7.

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 JR. Neuroethics and the possible types of moral enhancement. *AJOB Neuroscience*. 2012;3(4):3–14.
13. de Melo–Martin I, Salles A. Moral bioenhancement: Much ado about nothing? *Bioethics* 2014;9702:124–31; 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: Akabayashi, A ed. *The Future of Bioethics: International Dialogues*. Oxford: Oxford Scholarship Online; 2014.
16. Rogers EM. *Diffusion of Innovations, 4th Edition* [e-book]. London: Simon and Schuster; 2010.
17. See note 16, Rogers 2010.
18. See note 1, Van Lange et al. 2013
19. Griesinger DW, Livingston JW. Toward a model of interpersonal motivation in experimental games. *Behavioral Science* 1973;18(3):173–88.
20. See note 3, Murphy et al. 2011.
21. Poole ME, 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 AE, Spurk D. How do objective and subjective career success interrelate over time? *Journal of Occupational and Organizational Psychology* 2009;82(4):803–24.
22. Joireman JA, Shelley GP, Teta PD, Wilding J, Kuhlman DM. Computer simulation of social value orientation: Vitality, satisfaction, and emergent game structures. In: *Frontiers in Social Dilemmas Research*. Berlin, Heidelberg: Springer; 1996;289–310.
23. In some simulations, weights for MaxDiff, $|S-O|$, and MinDiff, $-|S+O|$, were included to model (anti)egalitarian orientations.
24. Extracted from 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 benefit in this case have a coordination problem that destabilizes the attractor.
26. Axelrod R, Hamilton WD. The evolution of cooperation. *Science*. 1981;211(4489):1390–6.
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 $a^{*\prime}$. These can be evaluated according to the *current* utility function U, and if $U(a^{*\prime},o)-U(a^*,o)>0$, it is rational for the agent to change its utilities to match U'.
28. In general, it is 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):621–41).
29. See note 12, Shook 2012.
30. Riis J, Simmons JP, Goodwin GP. 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;1251–88; Wike R, Holzwart K. Where trust is high, crime and corruption are low [Blog post]. *Pew Research Center* 2008. Available at http://www.pewglobal.org/2008/04/15/where-trust-is-high-crime-and-corruption-are-low/ (last accessed 10 Sept 2015).
32. Van Lange PA. 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;7(2):129–45; De Dreu CKW. Social conflict. *Current Sociology* 2013;61:696–713.

35. For example, Cardenas JC, Mantilla C. Between-group competition, intra-group cooperation and relative performance. *Frontiers in Behavioral Neuroscience* 2015;9:1–9; Puurtinen M, Mappes T. Between-group competition and human cooperation. *Proceeding Biological Sciences, The Royal Society* 2009;276:355–60; Burton–Chellew MN, Ross–Gillespie A, West SA. Cooperation in humans: competition between groups and proximate emotions. *Evolution and Human Behavior* 2010;31(2):104–8; Bornstein G. Winter E, Goren H. Experimental study of repeated team-games. *European Journal of Political Economy* 1996;12(4):629–39.
36. Sidanius J, Veniegas RC. Gender and race discrimination: The interactive nature of disadvantage. In: Oskamp S, ed. *Reducing prejudice and discrimination (The Claremont Symposium on Applied Social Psychology)*. Mahwah, NJ: Routledge; 2000;47–69.
37. Bowles S, Gintis H. *A Cooperative Species: Human Reciprocity and its Evolution*. Princeton: Princeton University Press; 2011.
38. Their model revealed that: (1) groups with non-parochial cooperators have a disadvantage over other groups and, therefore, would not have evolved in the first place, however; (2) groups with parochial cooperators, which 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 PJ. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* 2003;100(6):3531–5.
40. Bostrom N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press; 2014.

https://doi.org/10.1017/S0963180116001109 Published online by Cambridge University Press