Satisficing altruism vs. maximizing altruism: when doing good is more useful than doing best

Essay for Evolutionary Game Theory

University of Groningen

Siebe Rozendal

13-04-2018

4,007 words

I. An Explanandum for Indirect Reciprocity

People like doing good. Helping others gives actors a warm glow effect (Andreoni, 1990). However, for a long time evolutionary biologists have struggled to explain altruism towards strangers: kin selection was the main explanation for altruism (Hamilton, 1964). When someone shares your genes, it is in your genes' interest to help them. Later work showed that repeated interaction could give rise to cooperative behavior as well through so-called reciprocal altruism (Axelrod & Hamilton, 1981; Trivers, 1971). "I scratch your back if you scratch mine." As Skyrms (2014) argues, this mutual aid can explain many cases of altruism, especially if we abandon the assumption of random pairing. However, in this paper I shall use a more recent concept: *indirect reciprocity*. It captures the idea that you'll help someone even though they themselves cannot reciprocate, because it enhances your reputation. The better your reputation, the more likely it is that you will receive help from others in the future.

Although indirect reciprocity can explain why altruism emerges among people unable to reciprocate, it is unable to explain the current level of altruism in present-day societies: it falls shorts of the maximum possible amount advocated for by utilitarianism (Singer, 1972). This is a phenomenon that needs to be explained. I argue that in most communities reputation is not given on the basis of *how much* the potential recipient has helped, but rather *whether* the potential recipient has helped in a way that is emotionally pleasing to the contemplating donor. Because individuals adopt the strategy that will benefit themselves most in the long run, they will choose the act that will be most emotionally pleasing to observers (i.e. *satisficing altruism*). As a result, the whole community is worse off than they could have been: the best state is when everyone maximizes their benefit to others as long as the costs are bearable (i.e. *maximizing*

altruism). I modify the model of Nowak & Sigmund (1998) to incorporate a second cooperative option: bestowing a larger benefit to the recipient at a larger cost to oneself. I argue that if the two options are available, but not recognized as distinct by observers maximizing altruism is not a viable strategy. I briefly discuss the implications of this and discuss which conditions need to be met to give rise to a different norm.

II. How Indirect Reciprocity works

In models of indirect reciprocity it is made impossible for recipients to reciprocate donations to themselves. Instead actors are paired and assigned roles randomly; one (the donor) gets to opportunity to make a donation to the other (the recipient) at a cost to themselves. The recipient does nothing. The game is based on the Prisoner's Dilemma with two options: defect (give nothing to recipient, costs nothing to donor) and cooperate (give a benefit to recipient, has a cost to the donor smaller than the benefit to the recipient). After each interaction individuals are redistributed. After a number of interactions individuals replicate themselves on the basis of their collected payoffs. Their offspring will employ the same strategy and if the amount of offspring is determined by their collected payoffs (their fitness).

If this was all, cooperation would not take hold in the population. It would not be rational to benefit others with no benefit to oneself and there would be a population of defectors. But Nowak and Sigmund (1998) added two elements to the model to capture the concept of reputation: image scores and conditional strategies. An individual's image score is based on their previous behavior as a donor. The better behaved an individual is in the role of donor, the more likely they are to benefit in the role of recipient. This allows individuals to behave cooperatively to build a reputation that benefits themselves. Having a good reputation is beneficial when it is recognized and rewarded by others. This is modelled by conditional strategies: one can bestow the benefit on individuals with an image score s above a threshold k.

Strategies can be grouped into three distinct categories: always cooperate (Cooperators), always defect (Defectors), or cooperate conditionally (Discriminators). Nowak and Sigmund (1998) found that when mutation was possible Discrimination is not an evolutionary stable equilibrium. A population of Discriminators is hurt by Cooperators, because Cooperators allow Defectors to infiltrate. Another problem for Discriminators is that punishment is costly. Reputation is based on one's behavior: defecting against other defectors makes oneself a

defector. This makes simple Discriminators harsh on each other: punishers are punished in a vicious circle, leading to an all defect population. Ohtsuki and Iwasa (2004) found that the only evolutionary stable strategies distinguished justified punishment from unjustified punishment. If a recipient had previously defected against a defector (justified punishment), they would be rewarded. If they had defected against a cooperator (unjustified punishment), they would be punished.

There are a few other conditions that are necessary for cooperation to emerge. First and foremost, the benefit needs to outweigh the cost. The smaller the cost-benefit ratio the more likely cooperation is to emerge. Second, information about reputation must be sufficiently well-spread. In a model we can assume that individuals' reputations are common knowledge, but in reality this does not hold. Reputations are not public but private and different individuals have different information. The spread of information can be represented by q, the fraction of the population if which any given individual knows the reputation. Nowak and Sigmund show that cooperation through indirect reciprocity can only be stable if q is larger than the cost-benefit ratio.

III. Why reputation does not track amount of good done

Now we know how indirect reciprocity works we can try to explain something new: individuals in the twenty-first century are not pure altruists. For example, only 24.9% of US adults volunteered their time in 2015 ("Volunteering in the United States, 2015," 2016), and although US adults donate roughly $2,500/year on average, only 6% of donations are to foreign causes (*2013 Assessment of US Giving to International Causes*, 2014). Although these are considerable levels of altruism, they fall short of the utilitarian ideal to give until the costs become significant to oneself (Singer, 1972). Current estimates state that saving a life costs between $3,500 and $7,500 (GiveWell, 2017) and people spend many times this amount for recreational purposes over their own lifetimes. Furthermore, we even expect others to act mostly out of self-interest, suggesting there exists a social norm of self-interest (Miller, 1999), rather than a norm of (extreme) altruism.

Why is this the case? I argue that bestowing bigger benefits has bigger costs, but not bigger reputational effects for a multitude of reasons. Before going into those reasons, I first want to briefly argue against two other potential explanations.

Two potential explanations

Remember that in order for cooperative behavior to be stable q (the fraction that receives information about an actor's behavior) needs to be larger than the cost-benefit ratio. Therefore, one might explain the lack of cooperative behavior either by arguing that q is too small, or that c/b is too large. I believe neither explanation is satisfactory. Consider the first explanation; one might argue that during the twentieth century social networks have grown larger due to globalization and increasing complexity. This has made it more difficult to keep track of everyone's reputation, resulting in a low q (i.e. any given individual only knows the reputation of a small fraction of the total population). This explanatory is unsatisfactory for multiple reasons. First, globalization has not only increased the number of people each person interacts with, it has also increased information flow drastically (especially since the rise of the internet). Uncooperative behavior is easily visible online, and extremely uncooperative behavior is published widely in newspaper, hurting individuals' reputation permanently. Furthermore, population are not homogenous and individuals are not paired randomly; we form smaller social networks. One's behavior may not be visible to the whole world, but it is rather quickly communicated with the people one interacts with the most.

Another possible explanation is that individuals only want to increase their reputation up to a certain threshold. Once this threshold is reached, the marginal cost of increasing reputation is at least as high as the marginal benefit of increasing reputation. In fact, Nowak and Sigmund (1998) incorporated a threshold h that works like this in their model, such that a player cooperates if both the recipient have an image score of at least k, and they themselves have an image score below threshold h. I cannot exclude the possibility of such a threshold, but neither do Nowak and Sigmund provide empirical support for its existence. I believe that the threshold is partly a product of the model, which simplifies the donor's options to cooperate/defect, whereas in practice donors have the choice of how much to cooperate, and I would expect that higher recipients' reputations yield higher amounts of cooperation.[1]

Reputation does not track bestowed benefits

If reputation is highly useful, relatively public and the cost-benefit ratio of altruistic behavior is small, then why do we not see large amounts of altruism in easily observable behaviors such as donations and career choice in developed economies? I argue that bestowing bigger benefits

[1] Also note that in these models actors perform much more altruistically if error is introduced: one needs to uphold a reputation much higher than the threshold that others have to account for errors in communication.

has bigger costs, but not bigger reputational effects. To make a convincing argument, I first need to support my assertion that reputation does not (linearly) follow the bestowed benefits. After that, I will provide some possible explanations for this phenomenon.

It would be ideal to collect empirical data directly to answer the question of whether reputation follows bestowed benefits linearly. So far, we only have indirect support for this assertion. For example, several studies found that individuals being observed to make consequentialist decisions in moral dilemmas (and thus bestowing larger benefits on others) were perceived as less trustworthy and were therefore less preferred as social partners (Everett, Pizarro, & Crockett, 2016; Hughes, 2017). If we accept that trustworthiness is a component of reputation (it affects the likelihood of receiving benefits), consequentialist decisions are not the optimal way to gain reputation. Furthermore, reputation and punishment are related concepts: punishment is either the result of low reputation, or causes a lowered reputation. Experimental evidence shows that punishment is often administered retributively, and almost perfectly correlated with moral outrage (Greene, 2008). This indicates that emotional responses affect reputation more strongly than deliberative responses do.

There are many possible reasons for why reputation does not track amount of good done. In experiments, subjects have repeatedly demonstrated *insensitivity to scope* (Desvousges, 1992; Fetherstonhaugh, Slovic, Johnson, & Friedrich, 1997) valuing more lives saved little or no more than fewer lives saved. It appears that subjects judge the goodness of an act on the basis a mental representation of the consequences – 2,000 birds at an oil spill and 200,000 birds at an oil spill are both represented by the same mental image, evoking the same emotional response (Kahneman, D., Ritov, I., Schkade, D., Sherman, S. J., & Varian, 1999). Another reason is that it is difficult to evaluate the consequences of an act, such that a donation to one charity is seen as equal to a similar donation to a different charity, even if charities' effectiveness have large differences (Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J., Kahane, 2014; MacAskill, 2015). There is even a growing body of literature suggesting that doing good may result in a loss of reputation. When an observer feels threatened in their self-concept as a moral person, they may respond by putting down the altruist, called *do-gooder derogation* (Minson, J. A., & Monin, 2012).

Furthermore, the world is arranged in a way that affluent Westerners can do the most good by benefitting their outgroup. After all, their ingroup is by definition well off (at least financially). However, evolutionary pressures has created ingroup favoritism: benefiting strangers instead

of closed one's is looked down upon and judged as 'less moral' (Hughes, 2017). What's more, an altruist wants the benefits they donate to be beneficial to their potential benefactors; helping in one's community might not bring about the most good (MacAskill, 2015), but it will be more visible and emotionally compelling than a statistic of having saved a life in a far-away place.

In conclusion, reputation seems to be attributed not by means of deliberate evaluation of the consequences, but as an emotional response to the behavior. Therefore, there are pro tanto reasons to presume that the largest amounts of reputation are not gained through bestowing the most benefits on others. This is troublesome for those of us who want society to be maximally well-off. If individuals act to uphold a certain level of reputation and the maximizing way is costlier than the satisficing way, individuals will choose to satisfice. Next, I will modify the model of Nowak and Sigmund (1998) to represent the two cooperative options: high cost high benefit, low cost low benefit. Afterwards I will assess the implications and discuss some shortcomings.

IV. Modifying the model of indirect reciprocity

A model necessarily simplifies, and the indirect reciprocity model does what it is intended to do: provide an explanation how altruism can emerge without direct reciprocity or kin selection. However, it is currently unable to explain why there is not a higher level of altruism. I believe this can be done by a few simple modifications. In the original model, the donor has two options:

1. *Defect* (D): no cost to donor, no benefit to recipient, -1 to image score
2. *Cooperate* (C): cost to donor, benefit to recipient, +1 to image score, where benefit > cost

I propose to modify the options in the following way, such that reputation tracks *whether* there is a benefit given, not by *how large* the benefit is:

1. *Defect* (D): no cost to donor, no benefit to recipient, image score $s-1$
2. *Cooperate Satisficing* (C(Sat)): cost c to donor, benefit b to recipient, image score $s+1$, where $b > c$
3. *Cooperate Maximizing* (C(Max)): cost $2c$ to donor, benefit $2b$ to recipient, image score $s+2r$, where $b > c$, and $0 \leq r \leq 1$

When C(Max) is played, *r* is determined by the donor's contingent factors, such as their social network's sensitivity to the previously described psychological mechanisms and the nature of the cooperative act. If these are the options, only when *r* is on average large enough will C(Max) be evolutionary successful. If $\bar{r}$ = 0.5, strategies playing C(Sat) receive on expectation the same benefits as C(Max), but C(Max) pays higher costs. To successfully invade C(Sat), C(Max) must have an $\bar{r}$ large enough to generate future benefits to compensate for the extra cost *c* paid. This point is met if $\bar{r}$ raises the probability *(P)* of receiving an extra benefit *b* such that *P*b = c*. Where this point lies depends on the exact relationship between image score and the probability of (larger) future benefits, which would be needlessly complex to specify here. However, the more sensitive strategies are to the reputation of the potential recipient, the more beneficial it is to have a higher reputation. A more realistic (and inevitably more complex) model would have strategies with many thresholds *k*, such that different levels of reputation are rewarded (or punished) by different amount of benefits (or costs).

What we see is that it is theoretically possible to increase cooperation beyond Satisficing, but only if enough observers recognize the difference between Satisficing and Maximizing (i.e. $\bar{r}$ is large enough). It is not enough for actors to realize that have multiple cooperative options, but also that they are rewarded for maximizing altruism. I have intentionally not specified how *r* is determined. This will be sensitive to differences between cultures, social groups and individuals. If we relax the assumption that one's image score is set by random others we can instead allow for homophily (a network property that individuals associate and bond with similar others). This could create clusters of C(Max) players who each recognize each other's larger contributions and reward it appropriately (increasing $\bar{r}$ considerably).

V. Shortcomings

The proposed model modifications keep an assumption intact that is unlikely to be met in reality: that there is one population of donors and recipients. If there are two or more populations where only one population can play the donor role (e.g. well-off Westerners and poor developing world inhabitants), a donor population that strives to do maximally good cannot be rewarded for their altruism, because it would always be better to benefit the poorer outgroup. This is one explanation for ingroup favoritism and explains the 6% to foreign aid figure in section III; it is evolutionary beneficial. This is remarkably similar to the worry about whether consequentialists can be good friends (Mason, 1998; Railton, 1984). However, this worry can

be addressed from both a consequentialist and an evolutionary game theoretic standpoint. A consequentalist can argue that a supportive community of altruists will bring about much better consequences than an unsupportive one in the long run. Additionally, the model can be modified such that players can play *extra* games in which they can choose to benefit others from the same population only. This would reflect the fact that some types of 'expenditures' can only benefit those near to us: e.g. paying a compliment or offering emotional support.

Nonetheless, one may worry that I have evaded a crucial notion to altruism: the *cost/benefit-ratio.* After all, this ratio must be small enough for altruism to emerge. Although I have argued that we can save a life for roughly $3,500, this is not the benefit that matters for this ratio. The benefit that matters is the benefit that one stands to gain from altruism, not the benefit than one can bestow on others. In the end it is the benefit one gains that determine one's fitness. It may well be true that the utility loss from giving away $3,500 is not significantly compensated by reputational benefits. I think this is exactly the problem, but that we needed an explanation for *why* the reputational benefits do not compensate the costs.

I have excluded the possibility for r to be smaller than 0 for simplicity's sake. But we know that do-gooders may also be punished for doing more than the norm (Minson, J. A., & Monin, 2012; Stouten, van Dijke, Mayer, De Cremer, & Euwema, 2013). If we allow r to go negative, this gives C(Sat) players a tool to protect themselves against invasion: C(Max) mutants are easily eradicated. This means that punishment is possibly a double-edged sword for altruism: it allows protection against free-riders (Nowak & Sigmund, 2005), but can also be employed to maintain a level of altruism below the maximally possible level.

VI. Conclusion

I have argued that although the theory of indirect reciprocity is useful, it is as of yet unable to explain the limited levels of altruism in present-day society. If altruism gives much reputation and reputation is rewarded, then why do we not see more altruism? I presented a theoretical explanation: the amount of reputation obtained is not fully determined by the size of the benefits one bestows on others. Due to several mechanisms, reputational accruement is a more emotional case. For example, humans are not very sensitive to scope: a thousand lives are not given a thousand times more value than one life. Furthermore, we gain more reputation for

helping an ingroup member than helping an outgroup member, even if we can help many more outgroup members for the same cost as helping one ingroup member (MacAskill, 2015).

As a consequence, maximizing altruists are not proportionally rewarded – and sometimes even punished (Minson, J. A., & Monin, 2012) - for their greater positive impact. Satisficing altruists reap similar or better rewards with less effort. Unintentionally society arrived at a suboptimal condition. Can we improve this situation, and if so, how? Here I will give some tentative answers.

To improve the current situation the following three conditions must be satisfied.

1. There must be a state which is better (according to a particular axiology).
2. The better state must be at least asymptotically stable. That is, once reached, a few deviations cannot upset the current state, and these deviations are compensated for.[2]
3. The state must be reachable from the current state.

Few would argue that the current state of the world is perfect, thus the first condition is easily satisfied. The second condition also seems satisfied: we can imagine a possible world wherein reputation is given on the basis of good done, and playing any other strategy will yield lower expected payoffs. The third condition is more questionable: there must be some way for C(Max) players to survive and grow in number in the face of an overwhelming majority of other strategies, such as Defect and C(Sat). I believe this is possible and relates strongly to the second question: how can we improve the current situation?

Theoretically the advice is simple: do-gooders need to be rewarded on the basis of their effects, not on whether their actions are emotionally pleasing. In practice, this is very difficult. Humans are intuitive deontologists and operate on brains developed in small hunter-gatherer groups (Greene, 2008). Changing what yields reputations means changing what the members of a society value. Although philosophers such as Singer (1972) argue that people already value a human life in a way that should imply high amounts of altruism, few people agree with his conclusion and put that in practice. This article has offered an additional explanation why so few people implement Singer's conclusions: it is difficult to display high amounts of altruism if this is not recognized by others as better and thus not rewarded socially. It also shows a potential solution: clustering may provide the social rewards that are needed to maintain high

[2] This is related to Binmore's and Rawls's point that reasoning about the ideal society consists of choosing among different possible equilibria (Binmore, 2007).

levels of altruism. In such a cluster one receives reputation and corresponding benefits proportionally to the amount of good done.[3] If the members in the cluster successfully recognize and significantly reward amount of good done, they will each profit enough to persist and even grow.

Such a cluster has been created in the past decade. The organization Giving What We Can currently supports a community of over 3,000 people who have pledged to donate 10% of their annual income over their lifetime to the most effective causes. It has given rise to the effective altruism community, where allegedly 'doing the most good' is central (MacAskill, 2015) and rewarded socially and even financially (in terms of job offers for capable altruists). Although it is unlikely that this community will ever encompass a large share of the world population, some of its principles could be more widely adopted, such as giving more to charity, giving to the most effective charities[4], or focusing on unduly neglected causes. If these principles become widely valued, adhering to them will yield reputation and the associated benefits.

In conclusion, in order to increase the worldwide level of altruism, it is important to evaluate and reward altruism not only on the basis of what feels good, but also on the amount of benefits bestowed on others. Reputation is a powerful tool and should not be ignored by anyone who wants to improve the world.

References

2013 Assessment of US Giving to International Causes. (2014).

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464–477.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396.

Binmore, K. (2007). The Origins of Fair Play by Ken Binmore.

[3] Ideally individuals are rewarded for the *expected* amount of good done, since that brings about the best consequences over large numbers of actions and actors due to the Law of Large Numbers. This is even harder to evaluate than amount of good done and presents a considerable challenge to achieving the best state of the world, but I will leave it aside in this article.

[4] Many individuals care, or state to care, about the effectiveness of charities but often use imperfect metrics such as overhead (Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J., Kahane, 2014).

Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J., Kahane, G. (2014). The evaluability bias in charitable giving: Saving administration costs or saving lives? *Judgment and Decision Making*, *9*(4), 303.

Desvousges, W. H. (1992). *Measuring nonuse damages using contingent valuation: An experimental evaluation of accuracy*.

Everett, J. A. C., Pizarro, D., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General Everett Journal of Experimental Psychology: General*, *145*(6), 772–787. Retrieved from file:///C:/Users/mbesw/Downloads/SSRN-id2726330.pdf

Fetherstonhaugh, D., Slovic, P., Johnson, S., & Friedrich, J. (1997). Insensitivity to the Value of Human Life: A Study of Psychophysical Numbing. *Journal of Risk and Uncertainty*, *14*(3), 283–300.

GiveWell. (2017). *GiveWell's Cost-Effectiveness Analyses*. Retrieved from https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models

Greene, J. D. (2008). The Secret Joke of Kant ' s Soul. In *Moral Psychology* (pp. 35–117).

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, *7*(1), 17–52.

Hughes, J. S. (2017). In a moral dilemma, choose the one you love: Impartial actors are seen as less moral than partial ones. *British Journal of Social Psychology*, *56*(3), 561–577.

Humphrey, S. E., Nahrgang, J. D., & Morgeson, F. P. (2009). Integrating Motivational, Social, and Contextual Work Design Features: A Meta-Analytic Summary and Theoretical Extension of the Work Design Literature. *Journal of Applied Psychology*, *92*(5), 1332.

Kahneman, D., Ritov, I., Schkade, D., Sherman, S. J., & Varian, H. R. (1999). Economic preferences or attitude expressions?: An analysis of dollar responses to public issues. In *Elicitation of Preferences* (pp. 203–242).

Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences*, *107*(38),

16489–16493. http://doi.org/10.1073/pnas.1011492107

MacAskill, W. (2015). *Doing good better: Effective altruism and a radical new way to make a difference.* Guardian Faber Publishing.

Mason, E. (1998). Can an indirect consequentialist be a real friend? *Ethics*, *108*(2), 386–393. http://doi.org/10.1086/233810

Miller, D. T. (1999). The norm of self-interest. *American Psychologist*, *54*(12), 1054.

Minson, J. A., & Monin, B. (2012). Do-gooder derogation: Disparaging morally motivated minorities to defuse anticipated reproach. *Social Psychological and Personality Science*, *3*(2), 200–207.

Nowak, M. A., & Sigmund, K. (1998). Evolutionof indirect reciprocity by image scoring. *Nature*, *393*(June), 573–577.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298. http://doi.org/10.1038/nature04131

Ohtsuki, H., & Iwasa, Y. (2004). How should we define goodness? - Reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology*, *231*(1), 107–120. http://doi.org/10.1016/j.jtbi.2004.06.005

Railton, P. (1984). Alienation , Consequentialism , and the Demands of Morality. *Philosophy & Public Affairs*, *13*(2), 134–171.

Singer, P. (1972). Famine, Affluence, and Morality.

Skyrms, B. (2014). *Evolution of the Social Contract*.

Stouten, J., van Dijke, M., Mayer, D. M., De Cremer, D., & Euwema, M. C. (2013). Can a leader be seen as too ethical? The curvilinear effects of ethical leadership. *Leadership Quarterly*, *24*(5), 680–695. http://doi.org/10.1016/j.leaqua.2013.05.002

Trivers, R. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Volunteering in the United States, 2015. (2016). Retrieved April 12, 2018, from https://www.bls.gov/news.release/volun.nr0.htm