van Veelen, Matthijs

Working Paper

The evolution of morality

Tinbergen Institute Discussion Paper, No. TI 2020-063/I

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: van Veelen, Matthijs (2020) : The evolution of morality, Tinbergen Institute Discussion Paper, No. TI 2020-063/I, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/229683

TI 2020-063/I
Tinbergen Institute Discussion Paper

The evolution of morality

Matthijs van Veelen[1]

[1] University of Amsterdam

The evolution of morality

Matthijs van Veelen[1,2]

[1]University of Amsterdam, The Netherlands.
[2]Tinbergen Institute, The Netherlands.

17th September 2020

Abstract

Most of the literature on the evolution of human pro-sociality looks at reasons why evolution made us not play the Nash equilibrium in prisoners' dilemmas or public goods games. We suggest that in order to understand human morality, and human prosocial behaviour, we should look at reasons why evolution made us not play the subgame perfect Nash equilibrium in sequential games, such as the ultimatum game and the trust game. The "rationally irrational" behavior that can evolve in those games is a better match with actual human behaviour, including ingredients of morality such as honesty, responsibility, and sincerity, and also less nice properties, such as anger, as well as the incidence of conflict. Moreover, it can not only explain why humans have evolved to know wrong from right, but also why other animals, with similar population structures and similar rates of repetition, have not evolved the morality that humans have.

1 Nash equilibria and subgame perfection

The most important concept from game theory is the Nash equilibrium (Nash, 1950). A Nash equilibrium is a combination of behaviours, where none of the agents, or players, has an incentive to deviate. A classic example is the Nash equilibrium of the prisoners' dilemma. The prisoners' dilemma has two players, both of which can choose between cooperate, C, and defect, D. For both players, playing D is better than playing C, whether the other player cooperates or defects. In the example below, that is indeed the case; $1 > 0$, so defecting is better if the other one defects, and $3 > 2$, and hence defecting is also better if the other one cooperates. The Nash equilibrium therefore is for both to play D. What makes the prisoners' dilemma interesting, is that the Nash equilibrium does leave room for mutual improvement, as both would be better off if they both would cooperate instead of both defecting. In the example below, both could earn a payoff of 2 instead of 1 if they did.

$$\begin{bmatrix} & C & D \\ C & 2,2 & 0,3 \\ D & 3,0 & 1,1 \end{bmatrix}$$

The second most important concept from game theory is subgame perfection. This becomes relevant when the game consists of a sequence of moments in time, and players have to make choices at these different moments. This concept assumes that if the game is reduced to what remains of the game at a certain moment in time, given a combination of past behaviors of the players, then the players will play a Nash equilibrium in that remainder of the game, which is also called a "subgame". Moreover, this concept assumes that players correctly anticipate the future behavior of themselves and each other in the different scenarios that could unfold.

One classic example is the ultimatum game (Güth et al., 1982), which is played between a proposer and a responder. The proposer can propose a way to distribute a given amount of money, say 10 euros. The responder can then accept or reject that proposal, and in case she rejects, neither of the two gets any money. If the proposer proposed $10 - x$ for herself and x for the responder, then the responder chooses between, on the one hand, accepting and getting x, and, on the other hand, rejecting and getting 0 (here we for now only focus on how much the responder gets herself; later we will see how we may have

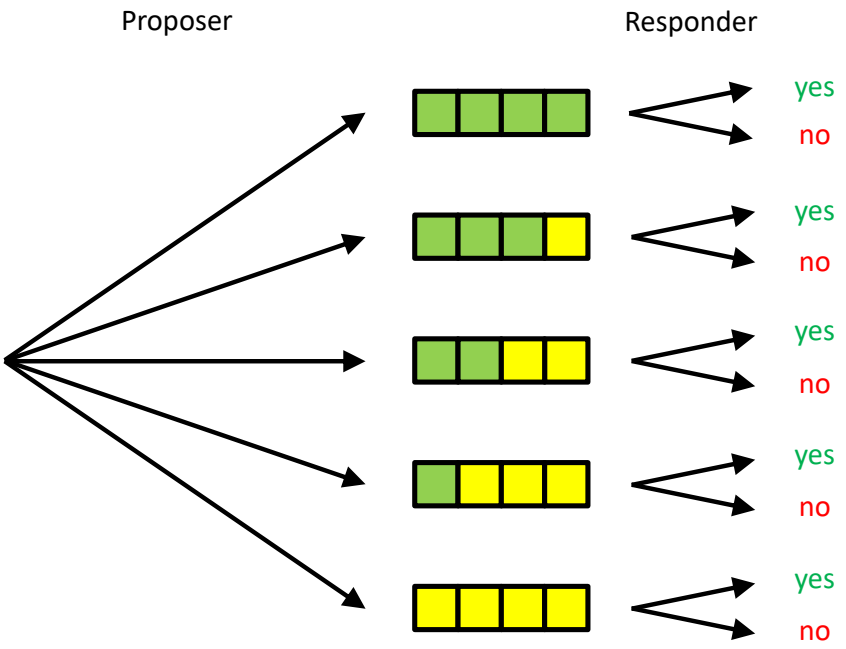

Figure 1: A simple version of the ultimatum game. The proposer chooses between proposals in which, from bottom to top, she gets 4, 3, 2, 1, and 0 herself, and the responder, also from bottom to top, gets 0, 1, 2, 3, and 4. For every proposal, the responder chooses whether or not to accept it. If the responder can commit to rejecting the bottom two proposals, the proposer is best off proposing an equal split.

evolved to not only look at our own payoffs). The Nash equilibrium at the second stage, where only the responder makes a choice, would therefore be for the responder to accept as long as $x > 0$, while she would be indifferent between accepting and rejecting when $x = 0$. If we now assume that x can only be an amount in whole euros, then there are two subgame perfect Nash equilibria in pure strategies. In the first, the responder accepts every possible proposal, and the proposer, anticipating that all proposals are accepted, proposes 10 for herself and 0 for the responder. In the second equilibrium, the responder accepts every proposal, except for the one in which she gets 0, which she rejects. The proposer anticipates that, and proposes 9 for herself and 1 for the responder. (Here we assume that players do not randomize. If we allow them to randomize, we would get more equilibria, but in none of those does the responder ever get

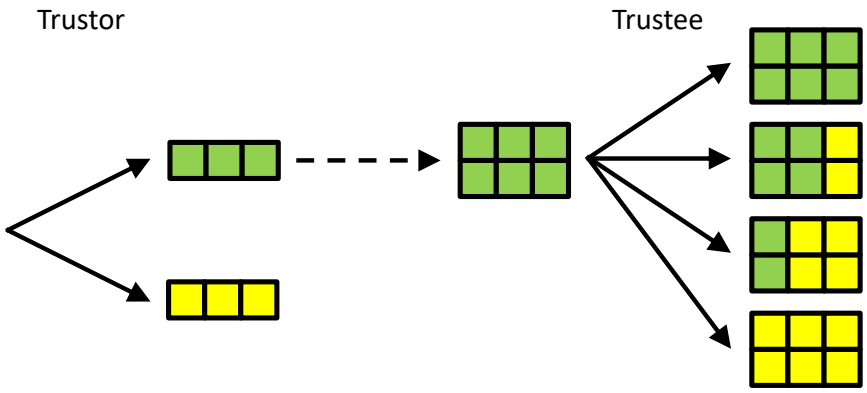

Figure 2: A simple version of the trust game. The trustor chooses whether or not to entrust the trustee with 3 euro's. These 3 euro's are doubled when entrusted to the trustee, who then gets to decide how much to send back; 0, 2, 4, or all 6 euro's, from top to bottom. If the Trustee can commit to sending back 4, the Trustor is best off entrusting the Trustee with the money. Compared to the subgame perfect Nash equilibrium with selfish preferences, in which the Trustee does not return any money, and the Trustee does not send any money, this will be better for both.

more than 1). The process of finding the subgame perfect Nash equilibria by starting at the end of the game, determining what equilibrium behavior would be once the players would get there, and then working towards the beginning, assuming that players correctly anticipate the behavior in later stages, is called backward induction. Figure 1 depicts this game, but for simplicity it has only 4 euros to be divided.

Another classic example is the trust game (Berg et al., 1995), which is played between a trustor and a trustee. In this game, the trustor can choose an amount of money to send to the trustee. Let's say this amount can be anything from 0 to 10 euro's, and we can call this amount x. The amount that the trustor sends to the trustee gets multiplied, say by 2, and then the trustee can choose an amount y which she sends back to the trustor. At the second stage, the trustee always keeps more money for herself if she sends back nothing, and hence she will choose $y = 0$ for every choice of x. The trustor, anticipating that the trustee will never send back anything, will not send any money to begin with,

and choose $x = 0$.

What makes this game interesting, is that, like the prisoners' dilemma, there are combinations of choices that would have left both players better off than in the subgame perfect Nash equilibrium; any choice $x > 0$, in combination with any choice of y in between x and $2x$ would give both players more money. Figure 2 depicts a simplified version of this game, where the trustor is restricted to only two values (it can only send all, which is 3 here, or nothing), and if it sends 3, which is then doubled, the trustee can only send back multiples of 2.

2 Why we do not always defect in prisoners' dilemma's

There are very many papers that look at why evolution would make individuals *not* play the Nash equilibrium in the prisoners' dilemma or the public goods game. The explanations can be classified in three broad categories; population structure, repetition, and partner choice.

Population structure encompasses any deviation from a setup in which individuals are matched randomly for playing a prisoners' dilemma or a public goods game, and also compete with each other globally. Such a deviation can for instance be that interactions happen locally on networks (Allen et al., 2017; Lieberman et al., 2005; Ohtsuki et al., 2006; Santos and Pacheco, 2005; Santos et al., 2008; Taylor et al., 2007), or within groups (Akdeniz and van Veelen, 2020; Luo, 2014; Simon et al., 2013; Traulsen and Nowak, 2006; Wilson and Wilson, 2007). In many such models, local dispersal causes neighbouring individuals, or individuals within the same group, to have an increased probability of being identical by descent. If individuals compete as locally as they have their opportunities for cooperation, then the cancellation effect prevents the evolution of cooperation (Taylor, 1992a;b; Wilson et al., 1992), but if they compete less locally, then cooperation can evolve. With assortment that is the result of identity by descent – which is typically, but not always the case in this category – one can also see this as kin selection operating. Kin selection through kin recognition, where relatives choose each other to play cooperate with, also falls under this category.

The second category of explanations is based on the fact that when interactions are not one-shot, but repeated, this changes the game, and allows for equilibria with cooperation. If the game is not over after the first prisoners'

dilemma, then there are opportunities for both to reward cooperative behavior, and retaliate against defection. There is an extensive literature on the large variety of equilibria that the "shadow of the future" creates (Fudenberg and Maskin, 1986), and their relative stability (Axelrod and Hamilton, 1981; Bendor and Swistak, 1995; García and van Veelen, 2016).

Partner choice is by far the smallest category. Here the idea is that if we can select who we play the game with, then we can select cooperative traits in each other. We will return to this category later.

These categories are very broad, but even then, the boundaries are not set in stone. Partner choice for instance can be seen as an endogenous source of assortment. Also some models combine ingredients from different categories, such as repetition and partner choice (Fujiwara-Greve and Okuno-Fujiwara, 2009), or repetition and population structure (Van Veelen et al., 2012).

3 Why we do not always play subgame perfect equilibria

The literature on why evolution would make individuals play strategies that are not subgame perfect is much smaller than the literature on the prisoners' dilemma. We do however believe that for understanding human pro-social behavior, understanding evolution in these games is at least as important. What these games have in common, is that players that are playing them would benefit from being able to commit themselves to a certain strategy. We can illustrate that with the ultimatum game.

If the proposer knows that the responder will accept anything, then the proposer will propose 10 for herself, and 0 for the responder. If the proposer knows that the responder will only accept offers in which she gets at least a certain minimum amount, say x, it will be in the proposer's own best interest to accommodate that, and propose $10 - x$ for herself and x for the responder. Therefore it would be advantageous for the responder if she were able to credibly commit to as high as possible a minimum amount she would accept. That is of course complicated by the fact that, once the proposal is on the table, what is in her best interest changes to just accepting anything positive. It seems however that evolution may have found a solution by endowing us with "rationally irrational" preferences, that make us prefer a situation in which we both end up getting nothing, if the alternative is that the proposer gets away

with proposing a (very) unequal split.

A similar commitment issue is central to the trust game. If the trustee would be able to commit to sending back an amount somewhere between the amount that the trustor will send her way, and twice that amount (which is how much there is after multiplication), then both would be the better for it; the trustor would gain by sending money and getting more money returned, and the trustee would have something instead of nothing. Again, committing is not easy, because once the money is sent and doubled, it is attractive for the trustee to keep everything. And, again, one possible interpretation of the empirical evidence is that evolution has found a solution by giving us "rationally irrational" preferences, that make us feel bad about keeping all the money (Johnson and Mislin, 2011).

The ability to commit can help an individual in two different ways. The first is that, when matched to a given partner, commitment can influence the behaviour of that partner. It does that in the ultimatum game, where committing to rejecting very asymmetric proposals can induce the proposer to make more generous proposals, and it does that in the trust game, where committing to sending back money can induce the trustor to send money. It is however also possible that who plays with whom is not fixed, and people can choose their partner. If there are two possible trustees, and one trustor, and one of the possible trustees has an irrational preference to send back a sizeable share, and the other does not, then the trustor would pick the irrational trustee, who then benefits from being picked. This version of the game one could call the who-to-trust game. Both with and without partner choice, this only works if players have a way of knowing what others are committed to.

The idea that the purpose of our moral sentiments is to allow us to credibly commit to certain, otherwise irrational behaviours is by no means new; it is the central premise of Frank (1987), as well as the book *Passions Within Reason*, also by Robert Frank (1988), who in turn refers to *The Strategy of Conflict* by Thomas Schelling (1960) as a source of inspiration (see also Schelling, 1978). But even though this idea is not at all new, most of the literature on the evolution of cooperation, human and non-human, is looking for reasons why evolution made us cooperate in prisoners' dilemmas or public goods games. In the remainder of this paper, we will go over reasons why evolution in games where commitment matters explains parts of human behaviour that evolution in the prisoners' dilemma or the public goods game cannot, including less rosy sides of our human nature.

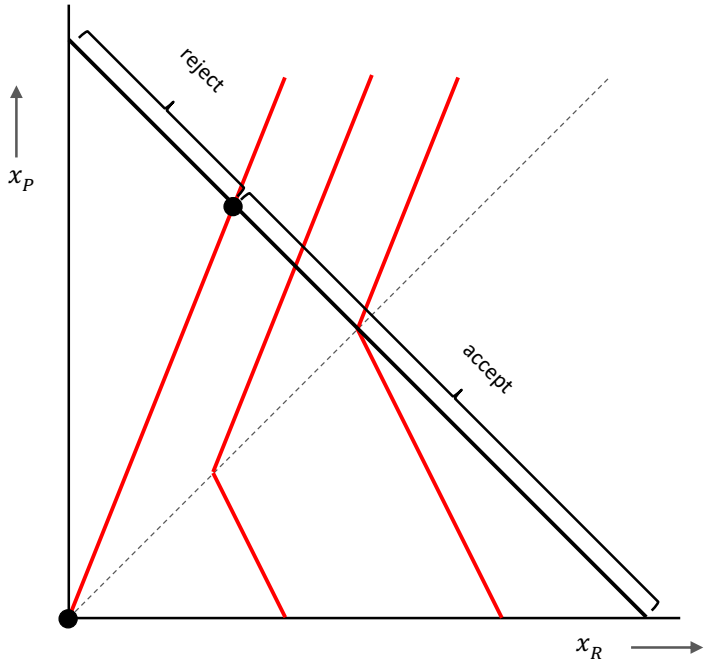

Figure 3: Indifference curves for a responder that has Fehr-Schmidt inequity averse preferences with $\alpha = \frac{2}{3}$ and $\beta = \frac{1}{3}$. The responder is indifferent between proposals (x_R, x) on one and the same red line, and likes proposals more to the right better than proposals more to the left. A proposer that also has Fehr-Schmidt inequity averse preferences would maximize his or her utility by choosing the point where the responder barely accepts (barely prefers the proposal over both getting 0), unless the proposer has an $\alpha > 1$, in which case she would propose an equal split.

4 Human behaviour

4.1 Ultimatum games and prisoners' dilemmas

One approach to understanding human behaviour in games like the ultimatum game is to assume that we have other-regarding preferences that evolved for other reasons, and that these preferences then also determine our behaviour in these games with backward induction (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999). These "other reasons" are for instance group selection or kin selection, and in models of group or kin selection, the game that is being played is typically either a prisoners' dilemma or a public goods game, or players have the opportunity to help each other in ways that add up to either of these. This

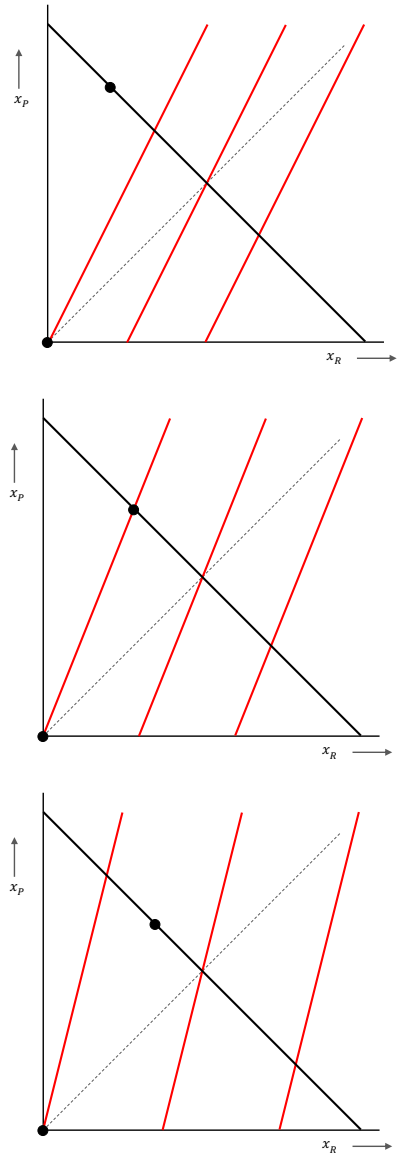

Figure 4: Here the preferences / the level of anger depends on the offer made (Cox et al., 2008; van Leeuwen et al., 2018); the responder would be sufficiently angry to reject the proposal in the first, barely accept it in the second, and accept it in the third subfigure.

approach with fixed other-regarding preferences is applied to the ultimatum game in Figure 3, where the preferences of the responder are treated as given,

and are independent of the proposal made. With Fehr-Schmidt inequity averse preferences, the responder dislikes being behind, and if she is behind too much, she will reject. The proposer will then propose a split that is barely accepted.

Another possibility is to assume that our preferences regarding other people can depend on what these other people do, and in this case that means that how the responder feels about the proposer, can depend on the proposal she makes (Cox et al., 2008). This is illustrated in Figure 4, where less generous offers lead to more anger and a higher willingness to forego money in order to punish the proposer. Moreover, we can assume that the way in which our preferences depend on other people's actions has evolved, not for getting the behaviour right in other games, like the prisoners' dilemma, but for getting the behaviour right in games like this, in which a commitment to reject low offers can help get better offers (van Leeuwen et al., 2018).

One empirical reason to believe that this second approach gives a better match with human behaviour, is that if the proposal is generated by a computer, responders do not reject quite as much as they do when the proposal is generated by the other person (Blount, 1995). Also, when an unequal split is proposed, and the only other option for the proposer was to propose an even more unequal split, the rejection rate is lower than when the unequal split is proposed, while the proposer could also have proposed an equal split (Falk et al., 2003). Both differences should not be there if the rejections were driven by proposals falling short of a fixed threshold for acceptance, as they are in the first approach. If the preferences have evolved to influence the behaviour of the proposer, on the other hand, they should be contingent on how much room to maneuver the proposer has.

A more theoretical reason to see a mismatch between current evolutionary (game) theory and the first approach concerns the jealous part of inequity averse preferences. This part reflects the dislike of disadvantageous inequality, which makes individuals willing to give up fitness in order to reduce the fitness of the other when the other is ahead and they themselves are behind. It is not clear how this would have evolved in a model in which players play prisoners' dilemmas or public goods games. In those games, players can choose to pay a cost in order to benefit other players, or they can choose not to, but they simply do not have the option to pay a cost to reduce the fitness of the other. In models with population structure where the option to pay a cost to harm the other does exist, spite can evolve as a consequence of negative relatedness (Hamilton, 1970), the same way altruism can evolve as a consequence of positive

relatedness (Hamilton, 1964a;b; van Veelen, 2007; van Veelen et al., 2017). In none of those models, however, do individuals evolve spite when behind and altruism when ahead at the same time.

Moreover, in the other direction, it seems that an element of commitment has actually snuck into our behaviour in prisoners' dilemmas and public goods games. Although some people are proper selfish and opportunistic, the majority are conditional cooperators in public goods games (Fischbacher et al., 2001) or prisoners' dilemmas (Charness et al., 2016); we are happy to cooperate, if the other one cooperates too, but if the other one defects, we prefer to defect as well. It seems therefore that evolution did not just make us indiscriminate cooperators or indiscriminate defectors – which is the menu of phenotypes in most models of evolution in the literature. Instead, evolution seems to have made us able to commit to not defect, as long as we are sufficiently sure that the other does not defect either. This can be interpreted as a spillover from repeated games, where reciprocal strategies can evolve, that stop cooperating if the other does not also cooperate. It is however important to realize that cooperation in prisoners' dilemmas can also evolve without repetition (or population structure). What is needed in this third scenario, is the ability to tell who is committed to cooperation, provided that the other one cooperates too, and common knowledge that both will cooperate. Knowing that the other will cooperate too is needed, because between two conditional cooperators, this becomes a coordination game with two equilibria; one where both play C; and one where both play D.

If conditional cooperators seek each other out for cooperation, then one mechanism at work would be partner choice, which would result in endogenous population structure. But also without partner choice, but with the ability to tell if others are also conditional cooperators, conditional cooperation can evolve. In this case, conditional cooperators would cooperate if they happen to be matched with each other, and defect if they meet defectors, and provided that they get it right sufficiently often what type the other is, that would give them a selective advantage.

4.2 Why we care for sincere altruism, true love, and good intentions

In order to illustrate why we would be preoccupied with sincerity, and value genuine caring more than opportunistic helping, we would like to introduce

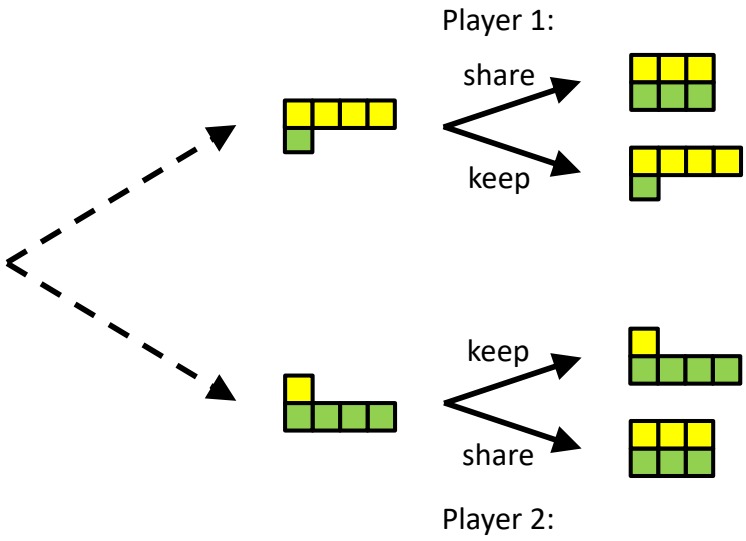

Figure 5: A simple version of the insurance game. Both players can be lucky or unlucky and the probabilities with which that happens are the same for both. If you are lucky, you have four, if you are unlucky you have one. If both are lucky, or both are unlucky (not depicted here), there is no use for helping. But if one is lucky, and the other is not, then helping will typically cost the lucky one less than it benefits the unlucky one. Ex post, after the dice are cast, it is better not to help, but if both would be able to commit to helping when the situation is uneven, this would, ex ante, be better for both.

another game, which one could call the "insurance game". In this game, there are two players that can either be lucky or unlucky. In this simple version, lucky means you get four, unlucky means you get one. If one is lucky, and the other one is not, then the lucky one can help the unlucky one, in which case both will end up with three. The idea behind this is that sharing is more beneficial for the unlucky one than it is costly for the lucky one.

In this game, it is a dominant strategy not to share, when you happen to be lucky, and the other one is not. However, if both players can commit to sharing, they would, in expectation, both be better off. If players that can commit would be able to single each other out, and play this game amongst themselves, they would do better than those that would never share and always keep what they

have.

In a population that is playing such a game, there could therefore be two related selection pressures. The first is a selection pressure to commit to sharing by genuinely caring for the other, which helps being chosen as a partner or friend. The second is a selection pressure to recognize genuine altruism, and distinguish it from fake displays of affection. Of course there is a tension that remains, as the best would of course be to be chosen as a partner or friend, be on the receiving end of sharing if you are unlucky yourself, and the other is not, but refuse to share when the tables are turned. But this tension is the whole reason why commitment would be needed in the first place, and it seems that the existence of sincere altruism and true love, as well as our preoccupation with distinguishing genuine care from opportunistic behaviour, indicates that evolution might have found a way to help us at least commit to a certain degree. It also makes sense that friendship and love typically converge to being symmetric partnerships, in the sense that people tend to end up being each others' friends, and if people stop liking us, we tend towards liking them less too.

Again, one could think of this as an extrapolation of reciprocity, which evolved in the context of repeated interaction, and there is of course no doubt that reciprocity has evolved in humans. But it is important to realize that we do not only pay people back, and say "you did the same for me", but also engage in hypothetical reciprocity, and say "you would have done the same for me" in cases in which we help a friend that has not had the opportunity to help us, and maybe never will. The latter would be consistent with the idea of evolved commitment in the insurance game, and that might be a better explanation than to consider this to be a maladaptive spillover from the reciprocity that has evolved for repeated prisoners' dilemmas.

If the insurance game is played repeatedly, and if helping a friend who is dealt a bad hand today increases her capacity for helping you in the future, then being committed to helping can also be in one's own self interest in a more direct way (Eshel and Shaked, 2001). Provided that both are committed to helping each other, then that help can be a great investment in receiving help in the future, not because you are investing in the other's *willingness* to help, as in standard models of reciprocity in repeated games, but in the other's *ability* to help, assuming the other's commitment is already there. A friend who you know would save your life, for instance, would not be around anymore to do that if you did not save hers, and hence it might be worthwhile taking a risk to do just that.

4.3 Why there is so much conflict in such a cooperative species

If the reason why humans are such a cooperative species is that evolution has favoured cooperation over defection because of population structure or repetition, then a reasonable question to ask is why there is not only cooperation, but also conflict and war. Cooperation in the prisoners' dilemma means paying a cost for the other to get a benefit. If we think of escalating a conflict as defection, and not escalating as cooperation, then it seems that in many situations, not escalating a conflict has a benefit-to-cost ratio that would be much more favourable for the evolution of cooperation than the benefit-to-cost ratio for many cooperative behaviours that humans engage in.

The presence of conflict can be explained by the fact that the evolution of commitment can be an arms race, and can result in populations in which some matches consist of individuals who have committed themselves to things that are incompatible. If both parties are committed to not budging – for which there may be good evolutionary reasons, because in many matches that results in a better outcome – then the result in this particular match is a conflict, where both now are worse off than they would have been, had they both been a bit more flexible and less committed.

A second remark to make here is that commitment not necessarily always serves the common good. It does create a win-win situation in the trust game and the insurance game, but in the ultimatum game, the money to be divided is a fixed amount, and all that commitment does, is help one party get a larger share of a pie of fixed size. In fact, commitment can also evolve in a version of the ultimatum game where the proposer can only offer for instance 75 cents on the dollar to the responder. In this case positive offers imply a decrease in efficiency. Most criminal activities, like for instance extortion, are efficiency reducing, and commitment also plays a large role there, as the ability to credibly commit to destructive action is a key component in the success of criminals.

4.4 Responsibility and honouring agreements

Many human collective action problems require some form of coordination. When people agree on doing a job together, and on a way to divide the different parts of that job, they all commit to doing their part, which becomes their responsibility. Given the agreement, not doing something that was your

responsibility will be frowned upon much more than not doing the same thing when it is not your responsibility. In other words, those that are involved will feel differently towards those that do not honour their agreements, and this can be seen as an (evolved) commitment to treating them less nicely than otherwise, or even punish them. One human solution to collective action problems therefore may have been to build a structure of commitments around them that can be turned on or off by mutual agreement. Human cooperation therefore is not necessarily just that we recognize public goods games, and play cooperate, but also that we have evolved to apply a flexible version of commitments to avoid all playing defect.

4.5 Altruistic 2nd party punishment

It has been widely recognized that punishment can sustain cooperation (Fehr and Gächter, 2002). This observation is typically followed by the realization that this is an incomplete explanation. While punishment may explain why there is cooperation, we would still need a reason why there is punishment, especially if punishment is costly (Brandt et al., 2006; Fehr and Gächter, 2002; Fowler, 2005; Hauert et al., 2007; Mathew and Boyd, 2009). One explanation for the existence of costly punishment is group selection. This is also a candidate to explain cooperation without the option to punish, but here it can be combined with the idea that, when established, punishment might be cheaper than the cooperation it enforces (Boyd et al., 2003). Higher order punishment might be even cheaper (Fehr and Fischbacher, 2003; Henrich and Boyd, 2001), but people do not really seem to use it (Kiyonari and Barclay, 2008). Another explanation would be the existence of the possibility to opt out of the public goods game, at a payoff that is higher than the payoff one gets if everyone defects. Models with this option predict cycles, and populations can spend sizable shares of their time in states where everyone cooperates and everyone would punish defectors (Brandt et al., 2006; Garcia and Traulsen, 2012; Hauert et al., 2007; Mathew and Boyd, 2009).

The premise of punishment as an incomplete explanation of cooperation, however, overlooks the possibility that even if punishment is costly, being committed to punishing may actually be beneficial, already for the individual, and the possible benefits to others might not be the reason why we punish, nor do we need the game to be voluntary. For making sure that we identify the possible advantages that commitment brings, it may perhaps be helpful to realize that a

prisoners' dilemma or public goods game with the option to punish is a proper different game than the prisoners' dilemma or the public goods game without punishment. With the option to punish, being committed to punishment might make the other players cooperate. If that happens often enough, then this can outweigh the costs of punishment when others defect, or the remaining deficit between individual costs and individual benefits may be so small, that it only takes a little bit of population structure to make the benefits to others outweigh that.

4.6 Recognizing commitment

What is needed for commitment to work, though, is that others can tell committed players apart. In the prisoners' dilemma or public goods game with punishment, being committed to punishment only has an advantage if others know. Similarly, being committed to rejecting unfair offers in the ultimatum game, or to returning money in the trust game, only works if others are aware of that. That implies that in experiments where subjects have no way to tell, or learn, who is committed and who is not, the individual benefit of being committed has no way of materializing.

In many experiments, subjects do not have the possibility to learn what other players are like individually, but there are some exceptions. In one experiment from Fehr and Fischbacher (2003), for instance, proposers in the ultimatum game get to see what the responder they are matched with accepted or rejected in past interactions with others. This not only allows proposers to find out what a responder will accept or reject, but it also opens the door for responders to strategically inflate their reputation for being a tough responder. That is also what happens; in the treatment with reputation, acceptance thresholds are higher. In the treatment without reputation, however, the acceptance threshold is not 0, as we also know from other experiments with ultimatum games. This is consistent with subjects being truly committed, and one could even say that trying to inflate your perceived level of commitment is only worth trying if there is also real commitment around.

This we expect to hold more generally. People cay be expected to understand the effect of being perceived as a committed punisher in games with punishment, as a compulsive money-back-sender in the trust game, and as a true friend in the insurance game. This would make them try to be seen as more committed than they really are. But even in settings where they cannot build a reputation,

if evolution made us able to truly commit, it will still show, whether the setting allows for the benefits of being committed to materialize or not.

Also in most models in evolutionary game theory, individuals typically adopt strategies that may depend on the other player's actions, but typically not on the other player's type. This is done with the idea of not imposing any form of rationality, which is a good point of departure, but it does rule out the evolution of commitment (see Alger and Weibull, 2012, for an exception).

4.7 Heterogeneity

For quite a few models of the evolution of cooperation, the form the prediction takes is that it separates parameter combinations where selection favours cooperation from parameter combinations where selection favours defection. In models that are not binary, but allow for a continuum of levels of cooperation, the prediction can be that there will be an equilibrium level of cooperation that everyone will settle on, which would also make for a homogeneous population (Allen et al., 2013). Another possibility is that evolution goes through a branching event, and settles on a stable coexistence of different levels of cooperation (Doebeli et al., 2004; van Veelen et al., 2017). This requires nonlinearity in the amount of cooperation. Other models that allow for the prediction to be a heterogeneous population are models with binary types, but with public goods games that are non-linear in the number of cooperators (Archetti, 2018; Archetti and Scheuring, 2012; 2016), dynamic greenbeard models (Jansen and Van Baalen, 2006) and some recent group selection models (Luo, 2014; Simon, 2010; Simon et al., 2013; van Veelen et al., 2014). For human behaviour, one might argue that the heterogeneity in cooperation is large enough to require an explanation (Andreoni and Miller, 2002; Fehr and Fischbacher, 2003), and that would speak for models in which heterogeneity can evolve.

What we would like to suggest, though, is that heterogeneity in human behaviour can perhaps also be described, not as a mix of cooperators and defectors, or as a blend of inequity averse and selfish individuals (Fehr and Schmidt, 1999), but as a combination of committed types and more opportunistic individuals. Not all models with commitment inevitably lead to heterogeneity, but some do. In the public goods game with punishment, the ability to commit can only make a difference if there are opportunistic others around, who will cooperate when they think they are matched with too many committed punishers. Opportunism on the other hand only pays if not everyone is (equally) committed to punish-

ment, and there is something to be opportunistic about. The presence of these types therefore only makes sense if they coexist.

4.8 From 2nd-party punishment to 3rd-party punishment and impartiality

The step from punishing those that hurt one's own interests to punishing those that hurt the interests of others is obviously non-trivial. One can however imagine that in between punishing someone in a 2-player game and impartial moral judgement concerning people other than yourself, there is an intermediate step, where individuals may benefit from committing to punishment in games with more than 2 players. The possibility to come to agreements and discuss what would and what would not be fair also seems to be a helpful step in paving the way for the concept of impartiality.

Both 2nd and 3rd party punishment in interactions that are not repeated are sometimes called altruistic. Punishing a defector after she defected on me might induce her to cooperate in later interactions with other individuals (Fehr and Fischbacher, 2003). This punishment then is thought to be beneficial to the next person she interacts with, and hence it is called altruistic. Also in 3rd party interactions, the idea is that those that benefit from the punishment are people that the wrongdoer will interact with in the future. When the mechanism behind the evolution of punishment is that commitment changes other people's behaviour, 2nd order punishment however does not have to be altruistic, because the real reason why one would be committed to punish could also be to not to be defected on oneself (see Section 4.5). Also with 3rd party punishment, the commitment might not be there to benefit the *next* person the wrongdoer meets, but to protect the *current* person she interacts with. In experiments where there is no way of learning whether someone is committed to punish, that might just fail to work, and only the collateral benefits to future interactants might show. This punishment therefore *is* designed to be altruistic, but not towards the next interaction partner, but the current one. This perspective is also more in line with the way in which Bernhard et al. (2006) find 3rd party punishment to be parochial. They conclude that the chances that an unfair choice by a 1st party is punished, are determined by whether or not the 2nd party belongs to the same group as the 3rd party, and not by the scope for bettering the behaviour of the 1st party in future in-group interactions, which would only be there if both the 1st and the 2nd party would belong to the same group.

5 Moral foundations

In his book *The Righteous Mind*, Jonathan Haidt (2012) describes five moral foundations.

- Care/Harm
- Fairness/Cheating
- Loyalty/Betrayal
- Authority/Subversion
- Sanctity/Degradation

These five dimensions of morality are meant as a starting point of a flexible description of the breadth of human morality.

If morality or pro-social behaviour would be limited to playing cooperate in prisoners' dilemmas or public goods games, then it seems we could probably make do with one dimension. If we consider the possibility that many ingredients of morality have evolved as a solution to a variety of commitment problems, then that still does not explain why Sanctity/Degradation would be a moral foundation, but it does allow for a richer description of morality. The role commitment plays in the insurance game, for instance, paints a more precise picture of the Care/Harm foundation, where sincerity matters. The flexible system of commitments that result from the possibility of making agreements and giving one's word, for example, would suggest a mechanism behind the Loyalty/Betrayal dimension.

Depending on one's taste in categorization, one could also argue that some solutions to commitment problems can be seen as moral foundations themselves. Honesty for instance is also a virtue (Purzycki et al., 2018), and although dishonesty can also be part of cheating or betrayal, it does not have to be. Being honest, or lying averse, is a commitment to reveal information that you have, and others don't. Being committed to truth telling can make one a more attractive partner in cases in which asymmetric information is likely to arise (Akdeniz et al., *in preparation*).

Also hypocrisy is a vice, and sensitivity to symmetry arguments, or being impartial, also when judging yourself, is a virtue. Since moral judgements are subject to debate, there is evolutionary pressure to be self-servingly biased (Babcock et al., 1995), and the more blatant forms of that are called out as hypocrisy.

6 Other species

If we consider evolutionary explanations for human morality, or human prosociality, then it is not only important that they give reasons why humans did evolve to be moral, but also why other species did not. The classical ingredients in explanations for the evolution of cooperation in prisoners' dilemmas and public goods games are population structure and repetition, and these two ingredients are indeed present in the human ecology. Humans are however not unique in living in (group) structured populations, nor are we special in interacting repeatedly. One way in which humans are at least somewhat special, is the way in which we make a living, and the incidence of commitment problems that that generates. That is not to say that there are no commitment problems elsewhere in nature, for which evolution may or may not have found solutions too, but at least compared to our close relatives, our niche is to acquire food in a way that requires more complex cooperation and more planning ahead. Human hunting for instance requires making tools in advance, and if I help you making your tool for our collective hunt, then I would really appreciate it if you don't subsequently trade it for something else. Also if you borrow a tool of mine, I would appreciate it if you would at some point return it. The technologically more elaborate, more information intensive way to make a living opens doors for opportunistic behaviour that remain closed in other species. If our morality is shaped to solve problems that do not exist in other species, or at least not to the same extent, then this also explains why we would be unique in our morality.

Some ingredients of morality moreover need language. Keeping your word, or committing to doing something by giving your word, means nothing without language. Honesty, being committed to tell the truth, obviously requires communication, and also hypocrisy is a discrepancy between words concerning others, and ones own behaviour, or the judgement that one expresses about it. This suggests that also language creates possibilities to commit in ways that are not available to other species.

References

Aslihan Akdeniz and Matthijs van Veelen. The cancellation effect at the group level. *Evolution*, 74(7):1246–1254, 2020.

Ingela Alger and Jörgen W Weibull. A generalization of Hamilton's rule—love others how much? *Journal of Theoretical Biology*, 299:42–54, 2012.

Benjamin Allen, Martin A. Nowak, and Ulf Dieckmann. Adaptive dynamics with interaction structure. *The American Naturalist*, 181(6):E139–E163, 2013.

Benjamin Allen, Gabor Lippner, Yu-Ting Chen, Babak Fotouhi, Naghmeh Momeni, Shing-Tung Yau, and Martin A Nowak. Evolutionary dynamics on any population structure. *Nature*, 544(7649):227–230, 2017.

James Andreoni and John Miller. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753, 2002.

Marco Archetti. How to analyze models of nonlinear public goods. *Games*, 9 (2):17, 2018.

Marco Archetti and István Scheuring. Game theory of public goods in one-shot social dilemmas without assortment. *Journal of Theoretical Biology*, 299:9–20, 2012.

Marco Archetti and István Scheuring. Evolution of optimal Hill coefficients in nonlinear public goods games. *Journal of Theoretical Biology*, 406:73–82, 2016.

Robert Axelrod and William Donald Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

Linda Babcock, George Loewenstein, Samuel Issacharoff, and Colin Camerer. Biased judgments of fairness in bargaining. *American Economic Review*, 85 (5):1337–1343, 1995.

Jonathan Bendor and Piotr Swistak. Types of evolutionary stability and the problem of cooperation. *Proceedings of the National Academy of Sciences*, 92 (8):3596–3600, 1995.

Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, 1995.

Helen Bernhard, Urs Fischbacher, and Ernst Fehr. Parochial altruism in humans. *Nature*, 442(7105):912–915, 2006.

Sally Blount. When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63 (2):131–144, 1995.

Gary E Bolton and Axel Ockenfels. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193, 2000.

Robert Boyd, Herbert Gintis, Samuel Bowles, and Peter J Richerson. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535, 2003.

Hannelore Brandt, Christoph Hauert, and Karl Sigmund. Punishing and abstaining for public goods. *Proceedings of the National Academy of Sciences*, 103(2):495–497, 2006.

Gary Charness, Luca Rigotti, and Aldo Rustichini. Social surplus determines cooperation rates in the one-shot prisoner's dilemma. *Games and Economic Behavior*, 100:113–124, 2016.

James C Cox, Daniel Friedman, and Vjollca Sadiraj. Revealed altruism. *Econometrica*, 76(1):31–69, 2008.

Michael Doebeli, Christoph Hauert, and Timothy Killingback. The evolutionary origin of cooperators and defectors. *Science*, 306(5697):859–862, 2004.

Ilan Eshel and Avner Shaked. Partnership. *Journal of Theoretical Biology*, 208 (4):457–474, 2001.

Armin Falk, Ernst Fehr, and Urs Fischbacher. On the nature of fair behavior. *Economic inquiry*, 41(1):20–26, 2003.

Ernst Fehr and Urs Fischbacher. The nature of human altruism. *Nature*, 425 (6960):785–791, 2003.

Ernst Fehr and Simon Gächter. Altruistic punishment in humans. *Nature*, 415 (6868):137–140, 2002.

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868, 1999.

Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics letters*, 71(3):397–404, 2001.

James H Fowler. Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences*, 102(19):7047–7049, 2005.

Robert H Frank. If *homo economicus* could choose his own utility function, would he want one with a conscience? *American Economic Review*, 77(4): 593–604, 1987.

Robert H Frank. *Passions Within Reason: The strategic role of the emotions.* WW Norton & Co, 1988.

Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

Takako Fujiwara-Greve and Masahiro Okuno-Fujiwara. Voluntarily separable repeated prisoner's dilemma. *The Review of Economic Studies*, 76(3):993–1021, 2009.

Julian Garcia and Arne Traulsen. Leaving the loners alone: Evolution of cooperation in the presence of antisocial punishment. *Journal of Theoretical Biology*, 307:168–173, 2012.

Julián García and Matthijs van Veelen. In and out of equilibrium I: Evolution of strategies in repeated games with discounting. *Journal of Economic Theory*, 161:161–189, 2016.

Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388, 1982.

Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion.* Vintage, 2012.

William D. Hamilton. The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):1–16, 1964a.

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

William D Hamilton. Selfish and spiteful behaviour in an evolutionary model. *Nature*, 228(5277):1218–1220, 1970.

Christoph Hauert, Arne Traulsen, Hannelore Brandt, Martin A Nowak, and Karl Sigmund. Via freedom to coercion: The emergence of costly punishment. *Science*, 316(5833):1905–1907, 2007.

Joseph Henrich and Robert Boyd. Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, 208(1):79–89, 2001.

Vincent AA Jansen and Minus Van Baalen. Altruism through beard chromodynamics. *Nature*, 440(7084):663–666, 2006.

Noel D Johnson and Alexandra A Mislin. Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889, 2011.

Toko Kiyonari and Pat Barclay. Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of personality and social psychology*, 95(4):826, 2008.

Erez Lieberman, Christoph Hauert, and Martin A. Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312, 2005.

Shishi Luo. A unifying framework reveals key properties of multilevel selection. *Journal of Theoretical Biology*, 341:41–52, 2014.

Sarah Mathew and Robert Boyd. When does optional participation allow the evolution of cooperation? *Proceedings of the Royal Society B: Biological Sciences*, 276(1659):1167–1174, 2009.

John F Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A. Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502, 2006.

Benjamin Grant Purzycki, Anne C Pisor, Coren Apicella, Quentin Atkinson, Emma Cohen, Joseph Henrich, Richard McElreath, Rita A McNamara, Ara Norenzayan, Aiyana K Willard, and Dimitris Xygalatas. The cognitive and cultural foundations of moral behavior. *Evolution and Human Behavior*, 39 (5):490–501, 2018.

Francisco C Santos and Jorge M Pacheco. Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, 95(9): 098104, 2005.

Francisco C Santos, Marta D Santos, and Jorge M Pacheco. Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454 (7201):213–216, 2008.

Thomas C Schelling. *The strategy of conflict*. Harvard University Press, 1960.

Thomas C Schelling. Altruism, meanness, and other potentially strategic behaviors. *American Economic Review*, 68(2):229–230, 1978.

Burton Simon. A dynamical model of two-level selection. *Evolutionary ecology research*, 12(5):555–588, 2010.

Burton Simon, Jeffrey A. Fletcher, and Michael Doebeli. Towards a general theory of group selection. *Evolution*, 67:1561–1572, 2013.

Peter D. Taylor. Altruism in viscous populations—an inclusive fitness model. *Evolutionary ecology*, 6(4):352–356, 1992a.

Peter D. Taylor. Inclusive fitness in a homogeneous environment. *Proceedings of the Royal Society B*, 249(1326):299–302, 1992b.

Peter D. Taylor, Troy Day, and Geoff Wild. Evolution of cooperation in a finite homogeneous graph. *Nature*, 447(7143):469–472, 2007.

Arne Traulsen and Martin A. Nowak. Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, 103(29):10952–10955, 2006.

Boris van Leeuwen, Charles N Noussair, Theo Offerman, Sigrid Suetens, Matthijs van Veelen, and Jeroen van de Ven. Predictably angry–Facial cues provide a credible signal of destructive behavior. *Management Science*, 64(7): 3364–3364, 2018.

Matthijs van Veelen. Hamilton's missing link. *Journal of Theoretical Biology*, 246(3):551–554, 2007.

Matthijs Van Veelen, Julián García, David G Rand, and Martin A Nowak. Direct reciprocity in structured populations. *Proceedings of the National Academy of Sciences*, 109(25):9929–9934, 2012.

Matthijs van Veelen, Shishi Luo, and Burton Simon. A simple model of group selection that cannot be analyzed with inclusive fitness. *Journal of Theoretical Biology*, 360:279–289, 2014.

Matthijs van Veelen, Benjamin Allen, Moshe Hoffman, Burton Simon, and Carl Veller. Hamilton's rule. *Journal of Theoretical Biology*, 414:176–230, 2017.

David Sloan Wilson and Edward Osborne Wilson. Rethinking the theoretical foundations of socio-biology. *Quarterly Review of Biology*, 82:327–348, 2007.

David Sloan Wilson, Gregory B. Pollock, and Lee A. Dugatkin. Can altruism evolve in purely viscous populations? *Evolutionary ecology*, 6(4):331–341, 1992.