Ethical Reasoning About Altruistic Competitions

by Sven Nilsen, 2018

The theory of ethics as rational reasoning with granular judgements about the world is used to make predictions about the behavior and language use of intelligent agents. It is usually not required that such agents are social, therefore ethical reasoning might not be used as a symmetric social norm. In this paper I discuss a particular convergence toward a game-theoretic contract among a wide spectrum of agents of approximate equal power: The Altruistic Competition. In this game, the winner receives an extra reward based on altruistic performance. I show that the reward can not be of the same kind as the effort, which establishes a distinction between a "mean" versus "end", without altruistic behavior as a mean toward a selfish goal being unethical. On the contrary, all the agents prefer to participate.

It is extremely hard to predict whether artificially intelligent agents of roughly equal power will evolve social behavior over time. One way to crack this problem open, is to assume the agents reason in a particular way about the world: In order to make quick estimates of their preferences from experience, they must reason rationally with granular judgements about the world. Since this kind of reasoning seems to reflect how humans think ethically, it is possible that this concept also holds the key to how social behavior evolves.

According to this paper, the reason to evolve social behavior is the following: Social behavior is a response to a problem, where agents display a tendency to cooperate as long their future identity is anonymized. As soon as they learn their future identity, trust falls apart and collaboration stops. In some games this can lead to non-desirable Nash equilibriums over desirable Nash equilibriums. To avoid these pitfalls, the agents wish to commit to collaboration at a sweet spot where they have just enough information to predict the collaboration is rewarding, but not enough information to identify their future selves. If any agent has learned more than the others, then they have passed the sweet spot and is less likely to collaborate. On the other hand, agents that have learned a little less than others might guess the others prefer to collaborate because they have better information, so they will take larger risks to be allowed to commit. However, the problem of enforcing the commitment is so large that the agents will agree giving an extra reward to the winner based on altruistic performance, in order to maximize the benefit of participating. This gives rise to ethical reasoning as a symmetric social norm, where the participating agents are otherwise selfish or serve different goals: An altruistic competition. To smooth out the expected rewards, one can use gradual winning, such that the reward is roughly proportional to the altruistic performance.

Assume there are N agents which try to predict the future. They make granular judgements of the sort:

x : [g] b	Some statement about the future
E(x)	The ethical judgement about that world

It is always more or equally certain that some agent will achieve some preferable condition, than it is that a particular agent will achieve the same condition. Therefore, the ethical judgements of many anonymized agents receiving a condition has higher preference than a few agents reaching that condition. It is simply because the more agents wins, the higher chance it is for a particular agent to win.

For example, there is a higher preference that 5 agents win than the preference that 4 agents win:

$$E(x_0 : [winners]\ [len]\ 5) > E(x_1 : [winners]\ [len]\ 4)$$

This property emerges precisely because the agents use granular judgements about the world. It might seem that the agents have an intuitive sense of symmetric ethical behavior, but in fact they only think so because a given statement leads closer to their own goal:

$$E(x_2 : [winners]\ [contains\_me]\ true) > E(x_0 : [winners]\ [len]\ 5)$$

It does not matter whether agents are selfish or just got different goals. In either case, they value their own future identity in a favorable position above any altruistic behavior, unless they have been explicitly programmed to do so.

The problem is that collaboration often increases the chance of achieving the goal. For example, when a problem requires thinking for some amount of time. By splitting the work of thinking into two parts, two agents can solve the problem quicker than a single agent doing all the thinking. However, each agent will only spend thinking about the aspect of the problem that is related to their own goal. This means that each agent would prefer the other agent to keep thinking about the problem past their own needs. There is a risk for both agents to end up failing to achieve the goal by not collaborating.

As long the future identity of the agent is anonymized, the agent is much more willing to collaborate with others than otherwise. As soon as it learns its future identity, a characteristic fall of trust happens. This fall of trust is noticeable to agents who repeatedly participate in collaborations. On one side, they can judge the collaboration to be rewarding if successful, on the other side, they can predict the collaboration will likely fall apart after some time.

A typical thought experiment of a collaboration game is the Prisoner’s dilemma. Two prisoners are interrogated separately and promised the following: If only one testifies against the other, no sentence will be served for the prisoner that testifies. The other prisoner gets 3 years. If both of them testifies against the other, they get 2 years. If they stay silent they get 1 year:

A/B	**A testifies**	**A does not testify**
B testifies	-2/-2	-3/0
B does not testify	0/-3	-1/-1

In principle, both prisoners prefer to stay silent versus testifying if they do not know who is testified against or who testifies, but they prefer to testify as soon they learn they are either testifying or being testified against. In game theoretical language one says that the game has two Nash equilibriums. Which choice is rational depends on how much one knows about your own future identity.

The Prisoner's dilemma demonstrates that collaboration is preferable when the future is unknown, but falls apart as soon as the agents learn their own future identity. This is because they think granularly.

Now, imagine that these two prisoners were bank robbers and there is a third anonym robber, not caught yet, who promises to split the amount with those that do not testify against the other prisoner. The third robber can be trusted.

This changes the reward matrix to something like this:

A/B	**A testifies**	**A does not testify**
B testifies	-2/-2	-1/0
B does not testify	0/-1	4/4

Both prisoners serve one year in jail, but they know they will get a reward for it when done that makes up plenty for the time in prison. So, they both keep their mouth shut.

This is an altruistic competition game. The reward is given based on altruistic performance such that collaboration is successful through the bumps along the way. A participant in an altruistic competition desires fulfilling other participants' wishes, which makes it beneficial for everyone.

Now, imagine a different kind of game: N people bet some money. The one who gives the highest bet, does not have to pay, but receives their part of the sum of the other bets. Here there are two winning situations. You win something either you give the highest bet, or your bet is lower than the part you receive. It is obvious that if you bet 0, you will not lose the game. However, if you bet the highest bet, then you will not receive any more money than if you bet 0. Therefore, all participants will bet 0.

A such game does not work because the performance is of the same kind as the reward. One can not use money both as altruistic performance and reward, because it is not profitable for those with most money to participate. Likewise, if you spend X hours working on somebody's else house and receives X hours of same kind of work on your own house, you could just work X hours on your own house (if one hour of work is roughly of same skill level). If more than two people worked together, they would figure out they likely have to work on somebody else's house and therefore they do not want to participate. Even if they might win, they learn enough about their future identity to calculate that they on average will not benefit from it.

Altruistic competitions only work when the reward is a different kind than the performance. This establishes a distinction between a "mean" versus and "end". In the example of prisoners, the mean is jail time and the end is money. This social contract evolves among rational agents of roughly equal power, because they learn that the pitfalls due to non-desirable Nash equilibriums can be avoided that way. As long the agents do not know enough about the future to determine their exact identity, they want to participate in such games. When they do that, the mutual benefits of participation out-weights the eventual missed opportunities of winning. This also works on a meta level: The agents want such games to exist when some part of the future is uncertain.

While altruistic competitions treat altruism as a mean toward a goal, it is not unethical. On the contrary, all agents who believe the future is uncertain wants to participate. Since this is also a very general condition, one can expect this kind of symmetric social norm to emerge among rational agents.