Possible Ways to Promote Compromise

Brian Tomasik
Center on Long-Term Risk
brian.tomasik@longtermrisk.org

Nov. 2013*

Abstract

Compromise has the potential to jointly benefit many different individuals, countries, and value systems. This piece enumerates ideas for how to encourage compromise, drawn from political science, international relations, sociology, and ethics.

Contents

*First written: fall 2013; last update: 5 Feb. 2016

1 Moral melting pot

One general approach is to bring together people of different moral views, so that they can sympathize with those who feel differently on moral issues. This helps promote tolerance of others, thereby improving odds for amicable resolution of disputes, and in some cases, each side may adopt some of the moral views of the other.

If we encourage people to move in each other's directions morally, is this actually compromise? Or are we just introducing a new morality that's a blend of the two? Well, consider the example of the deep ecologists vs. animal welfarists from the beginning of "Gains from Trade through Compromise." Suppose the deep ecologists and animal welfarists both look at the issue from the other side's perspective and thereby come to sympathize with it. Say the moral blend results in everyone caring half about deep ecology and half about animal welfare. Then the policies adopted when these morally blended individuals follow their own moral instincts will in fact be roughly the same as the compromise deals that would have been reached by the equipotent competing factions. So, even before the sides are morally blended, they should welcome an intervention in which someone convinces each side to move in the other's moral direction, so long as this moral blending is done roughly in proportion to the power of the existing sides.

What we see here illustrates a general principle: One function of emotions is as evolution's way of making game-theoretic pre-commitments. Romantic love is an emotional pre-commitment to provide care and resources for a partner. Anger is an emotional pre-commitment to respond against encroachment with costly retaliation. And, in this case, changing people's moral sentiments toward a compromise stance is a way to actually achieve lasting compromise. It may seem crude compared against carefully designed optimal game-theoretic bargains, but it has the advantage that it works now, without relying on institutional structures that can enforce contracts into the far future. And historical precedent shows us that modifying emotions for compromise purposes can work. For example, the advent of the value of religious tolerance may have been partly a response to the costly religious wars that plagued Europe. Of course, formal agreements like the Peace of Westphalia also contributed, and as is often the case, formal agreements can breed moral values just as moral values can lead to formal agreements.

I said above that moral blending is acceptable to both sides if the resulting blend is roughly proportional to the pre-existing power balance. However, sometimes this may not be the case. If it's not predictable which side people will favor upon considering both views, then ex ante, each side may still be okay with moral blending, because it's not clear which side will be favored more, and often, the members of each side think their stance is obviously more sensible, so they may even have high hopes for the outcome. Still, there are excep-

tions to this. For instance, members of certain religions and cults are discouraged from associating too closely with outsiders because this might predictably lead to straying from the fold. (2 Corinthians 6:14: "Do not be yoked together with unbelievers. For what do righteousness and wickedness have in common? Or what fellowship can light have with darkness?")

These are tricky cases, and there is a genuine tension between openness to new values versus goal preservation. For example, an altruist should rightly be concerned about marrying someone who spends all his money on expensive cars and tropical cruises, in part for fear of being tempted into the same lifestyle. That said, sometimes people are also flexible enough to "try on" other moral perspectives.

1.1 Meta-ethical views conducive to compromise

Beyond the game-theoretic reasons discussed above, there are at least three other motivations for openness to alternate moral perspectives:

1. Moral realism plus uncertainty: Often moral realists are ideological in defending the single position they think is "right," but reflective realists will notice that, just as with any other factual disagreement, we have significant uncertainty over what the moral truth is. Moreover, we should exercise epistemic modesty because other people may have information or insights we haven't yet discovered (Cowen & Hanson, 2002). Indeed, *if* the world consisted entirely of moral realists with common knowledge of each other's beliefs and modesty about where their priors came from (Hanson, 2006), then there would be no moral disagreement, at least in theory after enough computation (Aumann, 1976). There would still be moral uncertainty, but no one would have (altruistic) incentive to fight anyone else – only to learn from everyone else.
2. Moral non-realism plus extrapolation: Even those who don't believe in a single "moral truth" still generally feel that what they would want upon learning more, having a wider array of experiences, talking with more people, being less self-centered and more rational, etc. would be a better stance to take than what they want now. They would defer to this "extrapolated" or "idealized" version of themselves (Yudkowsky, 2004). This makes you more open to others' moral views because your own instantaneous introspection is imperfect, and other people provide a prior for the views that your future self might come to adopt. Other people feel as they do because of certain experiences and insights, and if you had those same experiences and insights, you might feel more the way they do. There are many different ways that extrapolation could be done, some more parochial than others, and depending on how much of your current self is preserved versus how much is allowed to float around, you'll end up with different levels of agreement. There's also not a unique convergent endpoint of extrapolation, because the process can be done in so many different ways, but some extrapolation procedures could enforce more convergence than others. Unlike with the previous two approaches, universal extrapolation wouldn't guarantee universal intergroup harmony, but depending on how much the algorithm is designed to enforce convergence, the result could still be a significant reduction in moral conflict.
3. Moral relativism: No moral view is more "right" than another. Everyone is entitled to her own view and should be allowed to act in accordance with it (maybe with some restrictions against major harm to

other cultures). As a result, moral relativists should not have inhibitions about trying on the views of others or reaching amicable agreements with others. If they were in charge of the world, relativists would not force anyone to follow their own personal moral views.

1.2 Is realism or anti-realism more favorable to moral convergence?

- Realism
  - *Anti-convergence*: Historically, moral realism has been a source of major conflicts, like religious persecution, ideological contests in politics, and other confrontations where each side thought it had "the truth" and was therefore justified in assaulting those who disagreed. Of course, realism has also been the dominant perspective historically, so we haven't had as much opportunity to observe what non-realism would have wrought.
  - *Pro-convergence*: A sophisticated perspective on moral uncertainty that takes peer disagreement seriously is driven to convergence because when others hold different views, this is evidence that you might be wrong and should update somewhat in their directions.
- Anti-realism
  - *Anti-convergence*: If moral attitudes are arbitrary, I may as well push for what I want. Why be more troubled by disagreement with other people than I am by disagreement with a pebble sorter?
  - *Pro-convergence*: If moral attitudes are arbitrary, I can see how what I want is due to specific patterns of neural wiring in myself, but they aren't somehow more sacred than the patterns of neural wiring in other people. We're all in the same kind of boat. So maybe I'll feel more sympathy for your neural wiring because, hey, I could have ended up with something like that too. Indeed, I may even extend nonzero sympathy to the pebble sorters.

Even if realism were better for convergence on balance, it's not clear we should encourage it wholesale, because it's somewhat confused. The idea of moral truth (whatever that's supposed to mean?) violates Occam's razor, and moreover, why would I care about what the moral truth was even if it existed? What if the moral truth commanded me to needlessly torture babies? It seems likely that as people become more sophisticated, they'll increasingly understand that naive realism doesn't make sense. Promoting it would then be like trying to tell kids about Santa to induce them to be nice rather than naughty; it only works for so long, especially among the really intelligent people who will have the most power over how the future unfolds.

That said, we may not want to promote concentrated anti-realism either, because this could encourage moral balkanization as people see that they can legitimately hold a stance in opposition to what others want. Rather, we should probably push most on convergence as a goal, like with coherent extrapolated volition, and not focus too much on the caustic nature of unadulterated anti-realism.

1.3 Other approaches

There are other ways to moral blending besides abstract meta-ethics. For example:

- Liberal education: Broadening people's horizons. Encouraging multiculturalism. Introducing people to diversity through the student body.
- Dialogue: Interfaith sharing of ideas. Events/books/TV shows where people of different views exchange perspectives.

Case studies of ideological disagreements and where people find common ground.

- Culture: Media that open people's minds and give them a sense of what it's like to be other people. Urban areas tend to be more cosmopolitan than rural ones, and insofar as this is caused by rather than merely correlated with geography, we could aim to replicate some of the same cultural dynamics more broadly.

Are these approaches cost-effective? Given that so many sectors of society aim to promote inter-group dialogue and reduce violence, we shouldn't expect low-hanging fruit here. On the other hand, because these efforts are widely regarded as beneficial, we have more confidence that working on them would at least be positive in expectation. Thus, these seem to be, at minimum, relatively "safe bets" and are supported by heuristics about working on causes that have wide support from many different people.

Keep in mind that some of the proposals, like promoting increased education, have many other flow-through effects that make the analysis more complicated. Promoting greater liberalism within existing education may be an easier intervention to analyze.

2 Us-vs.-them distinctions

Symbols, rituals, and shared identity can bind groups together, but usually this comes at the expense of increasing hostility toward outsiders. Oxytocin has this same effect. Jonathan Haidt suggests the metaphor that when people circle around a sacred object, they generate an "electric current" that unites them but also creates a polarity of them vis-a-vis the outgroup. Various studies have found that people identify with fellow group members more than outsiders even when the group assignments were basically arbitrary or even random. Even prelinguistic infants display this tendency, preferring puppets that have the same food preference – Cheerios or graham crackers – as they have (Neha Mahajana and Karen Wynn, 2012).

Ingroup loyalty can be strengthened by various factors, including anger ("Prejudice From Thin Air" by DeSteno, Dasgupta, Bartlett, and Cajdric (2004)) and the perception of zero-sum conflict (realistic conflict theory).

"Intergroup Conflict" has more to say, and the discipline of ingroup/outgroup distinctions has further literature on these topics (Hewstone & Greenland, 2000).

3 Transparency, social capital, and karma

Defection on one-shot prisoner's dilemmas happens because squealing on your partner doesn't have lasting consequences. If no one ever finds out, there's temptation to cheat. Real-world examples of this include lying, stealing, and other forms of deception for personal gain at greater social cost.

To prevent defection, it helps to make your choice visible to others, so that cooperation can be rewarded with future social benefits. One elegant way to accomplish this could be a "karma" rating or Whuffie score that is incremented or decremented based on how you treat others. If your fellow prisoner could lower your karma rating when you defect, you would have incentive not to do so.

As social networks become more ubiquitous, the possibility for these kinds of karma systems becomes more real. Already they exist in many online communities, like Slashdot or Quora. David Brin suggests that karma could become even more ubiquitous with digital glasses or other devices for looking up people's reputation scores in real-life settings.

Of course, similar incentive functions can also be accomplished with monetary payments, although it seems that there are social norms against paying money in certain circumstances, such as interactions between

friends. Instead, our more ancient primate sense of social capital still operates for relationships between family members, friends, business partners, and politicians. People feel less outrage about "bribery" when it's done through networking and social ingratiation rather than explicit financial exchanges.

All of these mechanisms facilitate compromise on prisoner's dilemmas by allowing for Pareto-improving transactions. As society becomes increasinglytransparent, it will be possible to enforce these arrangements in a greater number of cases, potentially making everyone better off, at least in theory. Already governments serve this function to a significant degree, by enforcing laws. (Of course, it's important to make sure that punishments for violations are roughly proportionate to the damage done rather than being excessive, or else these laws risk causing more harm than they prevent.)

Humans value privacy for many reasons, but one reason is because it historically offered protection – e.g., sneaking off to a secluded area to have sex so that the dominant male doesn't beat you up afterward. Similar principles apply in protecting citizens against authoritarian Big Brother. Avoiding authoritarianism is an important concern, but my sense is that this can be done by other means. For instance, Brin proposes sousveillance – watching the watchers to hold them accountable. As Brin says, safety is a necessary precondition for privacy, so at least some degree of surveillance is unavoidable.

Another reason we value privacy is because people tend to judge each other over trivialities – sexual conduct, religious beliefs, irreverent jokes, not being properly dressed, or whatever. The popularity of celebrity gossip and farcical political "scandals" are testaments to this feature of human nature. I think many people desire privacy because they don't want others seeing them doing these entirely normal activities that somehow are blown out of proportion when they're visible. If we could overcome the tendency to make a fuss over harmless private behaviors, it would allow for more transparency and therefore more opportunities for cooperation. As more of our lives become visible through digital technology, I hope people will become more accepting of diversity and individual choices, but getting there will sadly not be easy.

Transparency is probably even more important at a governmental level – the government both being transparent to its own citizens and being transparent to other governments. This can allow for better enforcement of international agreements, such as arms-control treaties (Shulman & Armstrong, 2009). In *The Strategy of Conflict*, Thomas Schelling (1980) explains (p. 148):

> Leó Szilárd has even pointed to the paradox that one might wish to confer immunity on foreign spies rather than subject them to prosecution, since they may be the only means by which the enemy can obtain persuasive evidence of the important truth that we are making no preparations for embarking on a surprise attack. [Citation: Szilárd's (1955) "Disarmament and the Problem of Peace"]

Transparency of citizens to governments is often protested, sometimes on privacy grounds and sometimes to prevent slipping down a slope toward tyranny. It's difficult to know where to draw the line on government surveillance. That said, it is clear that abuses of surveillance power – whether motivated by prurience or sabotaging one's opponents – are harmful, because they engender justified outrage at surveillance in general, making it slightly harder to carry out good surveillance.

One additional consideration is that surveillance and greater government power probably make eventual space colonization more likely by reducing catastrophic risks. This impact

may be met with ambivalence by those who consider preventing suffering most important.

4 Democracy, trade, and social stability

Democracy has many benefits for compromise.

1. Democracy itself seems to be among the best institutions devised to allow for positive-sum cooperation. Contrast it with monarchy or warlord rule, where people are violently overthrown, and the supporters of the overthrown party are tortured. Democracy allows for peaceful resolution of conflicts: If you don't like something, you don't grab a sword but instead grab a pen, or spend money, or otherwise exert power in a peaceful way. Legislative and electoral dynamics provide models for how compromises can be reached and maintained.
    - As discussed previously, (power-weighted) democracy should yield a Pareto improvement in ex-ante expected value to all parties. Even ex post, democracy is likely to be a Kaldor-Hicks improvement, because the citizens in the democracy will benefit more than the would-have-been autocrats lose.
2. Democracies also seem to lead to cooperation internationally. One explanation why is that democratic leaders have to prioritize public goods, including peace, more than authoritarian leaders do ("Game Theory, Political Economy, and the Evolving Study of War and Peace" by De Mesquita (2006), p. 639).
3. Finally, democracy has epistemic virtues as well because it aggregates information, insights, and analysis from many different sectors (Landemore, 2012). This allows people to find more ways to increase the pie, better insight into potential problems, and so on. In a speech on 24 Jan. 2002, Ralph Nader said, "Growing up civic is the liberation of the human mind and the facilitation of the greatest instrument ever devised to solve human problems, prevent injustices, foresee and forestall future perils and accentuate future benefits: a deliberative, working, daily democracy." (Nader, 2002)

Democracy relies on social stability. In order to be willing to compromise, you need to be confident the bargain will be upheld. Strong rule of law is required for this. In general, there is a vast literature on what makes democratic compromise possible, and we should explore it further. For instance, what helped trigger Democracy's Third Wave?

Trade tends to reduce the likelihood that factions or countries will go to war, because the parties rely on each other for mutual benefit. In addition, trade has a moral effect of enhancing empathy among distant peoples, as a natural corollary of the fact that reciprocal altruism leads us to care more about those with whom we exchange.

While this is the prevailing view among elites, there are some critics, such as Margaret MacMillan, who suggests that globalization can increase "intense localism and nativism," and this may have contributed to World War I; at the very least, growing interdependence didn't prevent that war. My personal guess, however, is that even if MacMillan's claim is true, it's a short-term effect, and the long-term trend is toward greater tolerance due to increased trade. In a session on "China Rising," Jon Huntsman gave as an example Utah's alfalfa exports to China as a factor helping to humanize and incentivize a more friendly relationship: The second largest economy in the world has gone "from enemy to customer" for those farmers.

5 Internationalism

Pride in one's country is a glue that holds countries together and justifies government build-up of force against other countries. John Mearsheimer said, "The most powerful political ideology on the face [of the planet ...] is not democracy; it's nationalism." And, he adds, nationalism makes it very hard to take over another country because the local population fights back unceasingly, as the US saw in Vietnam.

Of course, there are failed states, but the overall success of nationalism to promote unity even in spite of fierce ideological disputes is impressive. However, as is often remarked, the downside of nationalism is that it provokes *inter* national hostilities. This is the classic problem that Josh Greene (2013) discusses in *Moral Tribes*: the glue that turns "me" to "us" also pits "us" against "them."

The circle of what constitutes "us" can apparently grow large – from a 150-person tribal group to a 1-billion-person China, for instance. So it's not too much of an additional step to extend it to a 7-billion-person world. There is already an internationalist movement, aiming to encourage people to view each other as "citizens of the world." Or, in the words of John Lennon's "Imagine": "Imagine there's no countries / It isn't hard to do / Nothing to kill or die for / [...] Imagine all the people / Living life in peace..."

I don't know the cost-effectiveness of promoting internationalism relative to other things, but such interventions are at least very likely positive. Of course, if there are nearby extraterrestrials, the next steps will be interplanetism, intergalacticism, etc., but most people aren't ready for this yet. It would, of course, be a tragedy if internationalism led to fiercer conflicts with ETs, but hopefully the greater wisdom of our descendants will preclude that.

This area of ideas about how people perceive nationalism is one of the focuses of constructivists in international relations. Lisa Anderson's talk on "Nationalism and Ethnic Conflict" has further discussion about nationalism's origins and effects.

6 Cultural exchange

An important factor in cultivating internationalist sentiments is intermixing of people and cultures. For instance, one reason the US has had such strong ties with Europe is that many people of European descent live in the US, so that domestic political sentiments are aligned toward friendliness with European allies. This is even more prominently visible with American Jews exerting pressure for aggressive US backing of Israel, although in this particular case it's arguable that the domestic lobby causes more harm than good to international peace. Ideally, the cultural mix in the US would be sufficiently diverse that domestic politics wouldn't lopsidedly favor one foreign country over another.

Michael R. Auslin's (2011) *Pacific Cosmopolitans: A Cultural History of U.S.-Japan Relations* reviews a number of additional ways in which cultural exchange can improve international friendship, focusing on the case of Japan specifically:

- Cultural societies to spread the ideas of one nation to other nations
- The Fulbright Program and other exchange and study-abroad arrangements
- Trading appliances, music, food, sports players, etc.

Cultural icons like Nintendo from Japan or Jackie Chan from China are other examples of major forces that can help break down inter-country hostility in the minds of millions of ordinary citizens.

One concrete example where international exchange could be a cost-effective philanthropic project was described by George Perkovich in his interview with GiveWell. He

explained that in the 1990s, he helped with a program that brought together "young scholars and policymakers from Pakistan and India to increase goodwill and communication among the next generation of leaders" (p. 3). India and Pakistan opposed the program, and it shut down due to difficulty obtaining visas, but it could be resurrected. This seems important because the India-Pakistan conflict is arguably the most likely of any in the world to become nuclear.

6.1 Contact hypothesis

The sociology literature has extensively studied the contact hypothesis, which is the idea that people become more accepting toward those of different races, sexual orientations, or nationalities when they jointly have positive, coequal interactions in cooperative settings working toward common goals. This is what one would expect from the fact that positive-sum games require the brain to marshal warm feelings toward compatriots, to elicit cooperation rather than defection.

According to the the Wikipedia article, Donelson R. Forsyth's (2013) *Group Dynamics* meta-analyzed 515 studies and found a correlation coefficient of 0.2-0.3 between intergroup contact and absence of conflict. Thus, the hypothesis has extensive empirical backing, even though some researchers like Robert D. Putnam have found exceptions where more diverse communities display lower levels of trust.

As an aside, we might wonder whether the contact technique could be used to increase concern for animal suffering. In addition to having humans interact with animals directly, perhaps one could employ imagined contact or parasocial contact through the media, both of which have been suggested to help for human-human tolerance.

7 Female empowerment

Compared with men, women are generally less competitive and far less violent. One reason is that women are more risk-averse, because the number of offspring they can possibly have is bounded. Historically, men could acquire increased status and more sexual partners by succeeding in warfare against other tribes ("male warrior hypothesis"). Testosterone suppresses empathy and encourages conflict, "so much so that we had to invent sports to keep the boys happy,"notes Jonathan Haidt.

It thus seems plausible that as women gain more power in in a country, that country should ceteris paribus act more peacefully. David Pearce suggested female-only leadership as a way to reduce war, although it's unclear how big the effect would be, since structural factors might tend to select for the most competitive females to leadership roles. Also, some amount of willingness to fight can be important, for deterrence and humanitarian intervention.

Needless to say, Pearce's proposal would not be implemented any time soon. However, the more modest aim of empowering women seems valuable.

8 Global governance

Ultimately we might hope for the emergence of a world government, or singleton (Bostrom, 2006), which could provide the authority to enforce bargaining arrangements even at the international level. This idea is not new; compare with Hobbes's *Leviathan*, which argues for a sovereign authority to prevent "the war of all against all" (Hobbes, 1969). People give up their complete freedom in deference to a social contract that ultimately allows everyone to get more of what he wants in expectation than in a winner-take-all fight.

Short of a world government, we can aim for more modest forms of international coop-

eration. Many exchanges among nations can be seen as iterated prisoner's dilemmas, rather than one-shot versions of the game. As a result, neoliberal international-relations professor Robert Keohane has suggested the following ways to increase cooperation, as reported by Wikipedia's article on "Regime theory":

1. Make the rules and expectations clear, and require transparency and monitoring for compliance. Specify punishments for defection, such as sanctions.
2. Institutions can make it easier to negotiate on the margin once fixed costs are paid. For example, later rounds of the GATT agreements were facilitated by having already established negotiation processes in the earlier rounds.
3. Provide assurances that interactions will continue. The iterated prisoner's dilemma works best when agents foresee long futures of further engagement.

Finally, we can promote the idea of cooperation itself as the first resort to conflict, such as by fighting zero-sum thinking and explaining the logic of compromise. The school of liberalism in international relations takes a more positive view toward compromise and positive-sum possibilities than does realism, which tends to see one country's gain as another's loss. One reason for this is that realism tends to focus on *relative* gains, while liberalism emphasizes absolute welfare.

International conflicts have historically been among the most massive anthropogenic causes of death, so it seems that cooperation among nations (and major factions within nations) has high priority. The same may be true in an artificial general intelligence (AGI) race, if, for example, two major powers compete for control in analogy with the US and Soviet Union during the Cold War. It seems particularly important to raise awareness of these issues among potential AGI designers themselves, as well as the military and corporate leaders funding AGI projects, although general public outreach can still be valuable to the extent it influences these leaders by diffusion. (For example, the technologists who actually end up building AGI may be born 50 years from now and be influenced by parents, teachers, and TV programs that were in turn influenced by what we did today.)

9 Advancing compromise theory

Research on the fundamentals of compromise might have high payoff. There are still many issues in game theory that remain not well understood, and putting compromise on firmer ground would be a major step forward. In addition, we need research in political and social theory to devise robust mechanisms for sustaining compromise agreements, especially against potential disruptions to existing institutions that may result from fast technological breakthroughs or other black swans.

10 Improving rationality

Helping groups that are fighting over soluble factual questions might reduce many short-term conflicts. For more, see this discussion of epistemic disagreements.

11 Improved information?

Epistemic disagreements represent (theoretically irrational) divergences of opinion in the case of common knowledge. However, very often agents don't have the same knowledge. Games of imperfect information are often more prone to conflict than those of perfect information; under perfect information, the best outcome is usually to compromise. (Of course, there may be exceptions.) In other words, shifting situations from games of imperfect to games of (more) perfect information could be valuable. A downside is that more knowledge *in general* also means that risks come faster and may allow for less time in which to negoti-

ate and work out social structures that better facilitate a good outcome for everyone.

12 Compromise technologies?

While it's extremely important to promote cooperation, this field is not laden with low-hanging fruits, because many other people already rightly see its value. Historically, one major source of leveraged social change has been technologies that open up new possibilities. Are there technologies we could support that hold the promise of dramatically improving cooperation, without also speeding up dangers of massive conflict at the same time?

One proposal that a friend of mine suggested is improving machine translation. Language plays a major role in the development of national identities and us-versus-them balkanization. One of the goals of Esperanto was to "transcend nationality and foster peace and international understanding between people with different languages." While Esperanto has little hope of worldwide adoption, very readable machine translation would offer something almost as good. Apparently this vision has been floating around since the end of World War II.

13 What's the net impact of game theory?

In *Game Theory: Analysis of Conflict*, Roger B. Myerson (2013) suggests (pp. 1-2):

> People seem to have learned more about how to design physical systems for exploiting radioactive materials than about how to create social systems for moderating human behavior in conflict. Thus, it may be natural to hope that advances in the most fundamental and theoretical branches of the social sciences might be able to provide the understanding that we need to match our great advances in the physical sciences.

Of course Myerson is likely to feel this way because (a) if he didn't, he might not have studied game theory and (b) he probably doesn't want to feel as though his life's work has been harmful. But is it true that our prospects for reducing suffering are better when people are more informed about game theory?

It's certainly the case that there are both gains and losses when people understand game theory relative to relying on naive intuition or happenstance. Some examples of downsides:

- Standard game theory says it's rational to defect on a one-shot prisoner's dilemma, whereas some people intuitively would cooperate.
- As Josh Greene has noted, people intuitively cooperate on the public-goods game (a multi-player prisoner's dilemma) but stop doing so when they're in a more calculating mindset.

On the other hand, there are benefits to deeper understanding as well:

- Thinking about the inefficiency of war or litigation or attritional conflicts can help us
  see that, while sometimes these may be rational undertakings, sometimes they're not, and we may be able to avoid costly expenditures resulting from mistakes.
- Knowing how games work allows us to create mechanisms to prevent undesirable outcomes, such as by building side-payments into a one-shot prisoner's dilemma to enforce cooperation.
- In general, we may be able to modify the payoffs, including by changing people's attitudes, so that the situation is transformed to a different game that has a happier Nash equilibrium.

Game theory is destiny. In the long run, rational agents will converge on understanding game theory anyway, because those who don't will on average lose resources. If we can emphasize the positive possibilities of game the-

ory, we may be able to steer society toward a better path. Of course, *if*, hypothetically, it were the case that game theory tended to produce worse results the more it was understood, we might hope to keep people in the dark as long as possible in order to maximize cooperation before crucial junctures like the creation of AGI. However, I think this is not that likely, and indeed, the opposite seems more plausible: that better understanding of game theory would help navigate cooperation on AGI to make everyone better off in expectation.

14 Charities that promote cooperation

I have a separate piece that lists organizations that work to promote cooperation: "Cooperation Charities and Organizations."

References

Aumann, R. J. (1976). Agreeing to disagree. *The annals of statistics*, 1236–1239.

Auslin, M. R. (2011). *Pacific Cosmopolitans: A cultural history of U.S.-Japan relations*. Cambridge, Mass: Harvard University Press.

Bostrom, N. (2006). What is a singleton? *Linguistic and Philosophical Investigations*, *5*(2), 48–54.

Cowen, T., & Hanson, R. (2002). Are disagreements honest. *Journal of Economic Methodology*.

De Mesquita, B. B. (2006). Game theory, political economy, and the evolving study of war and peace. *American Political Science Review*, *100*(04), 637–642.

DeSteno, D., Dasgupta, N., Bartlett, M. Y., & Cajdric, A. (2004). Prejudice from thin air: The effect of emotion on automatic intergroup attitudes. *Psychological Science*, *15*(5), 319–324. doi: 10.1111/j.0956-7976.2004.00676.x

Forsyth, D. R. (2013). *Group dynamics* (6edition ed.). Belmont, CA: Cengage Learning.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them* (1st edition ed.). New York: Penguin Press.

Hanson, R. (2006). Uncommon priors require origin disputes. *Theory and decision*, *61*(4), 319–328.

Hewstone, M., & Greenland, K. (2000). Intergroup conflict. *International Journal of Psychology*, *35*(2), 136–144.

Hobbes, T. (1969). *Leviathan, 1651*. Scolar Press.

Landemore, H. (2012). *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton ; Oxford: Princeton University Press.

Mahajan, N., & Wynn, K. (2012). Origins of "Us" versus "Them": Prelinguistic infants prefer similar others. *Cognition*, *124*(2), 227–233. doi: 10.1016/j.cognition.2012.05.003

Myerson, R. B. (2013). *Game theory*. Harvard university press.

Nader, R. (2002). *Crashing the party: Taking on the corporate government in an age of surrender. new york: St.* Martin's Press.

Schelling, T. C. (1980). *The Strategy of Conflict*. Harvard University Press. (Google-Books-ID: 7RkL4Z8Yg5AC)

Shulman, C., & Armstrong, S. (2009). Arms control and intelligence explosions. In *7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July* (pp. 2–4).

Szilard, L. (1955). Disarmament and the Problem of Peace. *Bulletin of the Atomic Scientists*, *11*(8), 297–307.

Yudkowsky, E. (2004). Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*.