Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda

Sections 1 & 2: Introduction, Strategy and Governance

only read introduction

by **JesseClifton** 17th Dec 2019

Center on Long-Term Risk (CLR) | Risks of Astronomical Suffering (S-risks) | Research Agendas | Game Theory | Coordination / Cooperation | AI | Frontpage

This post is part of the sequence version of the Effective Altruism Foundation's research agenda on Cooperation, Conflict, and Transformative Artificial Intelligence.

1 Introduction

Transformative artificial intelligence (TAI) may be a key factor in the long-run trajectory of civilization. A growing interdisciplinary community has begun to study how the development of TAI can be made safe and beneficial to sentient life (Bostrom 2014; Russell et al., 2015; OpenAI, 2018; Ortega and Maini, 2018; Dafoe, 2018). We present a research agenda for advancing a critical component of this effort: preventing catastrophic failures of cooperation among TAI systems. By *cooperation failures* we refer to a broad class of potentially-catastrophic inefficiencies in interactions among TAI-enabled actors. These include destructive conflict; coercion; and social dilemmas (Kollock, 1998; Macy and Flache, 2002) which destroy value over extended periods of time. We introduce cooperation failures at greater length in Section 1.1.

Karnofsky (2016) defines TAI as "AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution". Such systems range from the unified, agent-like systems which are the focus of, e.g., Yudkowsky (2013) and Bostrom (2014), to the "comprehensive AI services" envisioned by Drexler (2019), in which humans are assisted by an array of powerful domain-specific AI tools. In our view, the potential consequences of such technology are enough to motivate research into mitigating risks today, despite considerable uncertainty about the timeline to TAI (Grace et al., 2018) and nature of TAI development. Given these uncertainties, we will often discuss "cooperation failures" in fairly abstract terms and focus on questions relevant to a wide range of potential modes of interaction between AI systems. Much of our discussion will pertain to powerful agent-like systems, with general capabilities and expansive goals. But whereas the scenarios that concern

much of the existing long-term-focused AI safety research involve agent-like systems, an important feature of catastrophic cooperation failures is that they may also occur among human actors assisted by narrow-but-powerful AI tools.

Cooperation has long been studied in many fields: political theory, economics, game theory, psychology, evolutionary biology, multi-agent systems, and so on. But TAI is likely to present unprecedented challenges and opportunities arising from interactions between powerful actors. The size of losses from bargaining inefficiencies may massively increase with the capabilities of the actors involved. Moreover, features of machine intelligence may lead to qualitative changes in the nature of multi-agent systems. These include changes in:

1. the ability to make credible commitments;
2. the ability to self-modify (Omohundro, 2007; Everitt et al., 2016) or otherwise create successor agents;
3. the ability to model other agents.

These changes call for the development of new conceptual tools, building on and modifying the many relevant literatures which have studied cooperation among humans and human societies.

1.1 Cooperation failure: models and examples

Many of the cooperation failures in which we are interested can be understood as *mutual defection in a social dilemma*. Informally, a social dilemma is a game in which everyone is better off if everyone cooperates, yet individual rationality may lead to defection. Formally, following Macy and Flache (2002), we will say that a two-player normal-form game with payoffs denoted as in Table 1 is a social dilemma if the payoffs satisfy these criteria:

- $R > P$ (Mutual cooperation is better than mutual defection);
- $R > S$ (Mutual cooperation is better than cooperating while your counterpart defects);
- $2R > T + S$ (Mutual cooperation is better than randomizing between cooperation and defection);
- For quantities $\text{greed} = T - R$ and $\text{fear} = P - S$, the payoffs satisfy $\text{greed} > 0$ or $\text{fear} > 0$.

		Player 2	
		Action 1	Action 2
Player 1	Action 1	R, R	S, T
	Action 2	T, S	P, P

Generic symmetric game

	C	D
C	$-1, -1$	$-3, 0$
D	$0, -3$	$-2, -2$

Prisoner's Dilemma

	C	D
C	$0, 0$	$-1, 1$
D	$1, -1$	$-10, -10$

Chicken

	C	D
C	$3, 3$	$0, 2$
D	$2, 0$	$1, 1$

Stag Hunt

Table 1: A symmetric normal-form game (top) and three classic social dilemmas (bottom).

Nash equilibrium (i.e., a choice of strategy by each player such that no player can benefit from unilaterally deviating) has been used to analyze failures of cooperation in social dilemmas. In the Prisoner's Dilemma (PD), the unique Nash equilibrium is mutual defection. In Stag Hunt, there is a cooperative equilibrium which requires agents to coordinate, and a defecting equilibrium which does not. In Chicken, there are two pure-strategy Nash equilibria (Player 1 plays D while Player 2 plays C, and vice versa) as well as an equilibrium in which players independently randomize between C and D. The mixed strategy equilibrium or uncoordinated equilibrium selection may therefore result in a crash (i.e., mutual defection).

Social dilemmas have been used to model cooperation failures in international politics; Snyder (1971) reviews applications of PD and Chicken, and Jervis (1978) discusses each of the classic social dilemmas in his influential treatment of the security dilemma.[1] Among the most prominent examples is the model of arms races as a PD: both players build up arms (defect) despite the fact that disarmament (cooperation) is mutually beneficial, as neither wants to be the party who disarms while their counterpart builds up. Social dilemmas have likewise been applied to a number of collective action problems, such as use of a common resource (cf. the famous "tragedy of the commons" (Hardin, 1968; Perolat et al., 2017)) and pollution. See Dawes (1980) for a review focusing on such cases.

Many interactions are not adequately modeled by simple games like those in Table 1. For instance, states facing the prospect of military conflict have *incomplete information*. That is, each party has private information about the costs and benefits of conflict, their military strength, and so on. They also have the opportunity to negotiate over extended periods; to monitor one another's activities to some extent; and so on. The literature on bargaining

models of war (or "crisis bargaining") is a source of more complex analyses (e.g., Powell 2002; Kydd 2003; Powell 2006; Smith and Stam 2004; Feyand Ramsay 2007, 2011; Kydd 2010). In a classic article from this literature, Fearon (1995) defends three now-standard hypotheses as the most plausible explanations for why rational agents would go to war:

- Credibility: The agents cannot credibly commit to the terms of a peaceful settlement;
- Incomplete information: The agents have differing private information related to their chances of winning a conflict, and incentives to misrepresent that information (see Sanchez-Pages (2012) for a review of the literature on bargaining and conflict under incomplete information);
- Indivisible stakes: Conflict cannot be resolved by dividing the stakes, side payments, etc.

Another example of potentially disastrous cooperation failure is *extortion* (and other compellent threats), and the execution of such threats by powerful agents. In addition to threats being harmful to their target, the execution of threats seems to constitute an inefficiency: much like going to war, threateners face the direct costs of causing harm, and in some cases, risks from retaliation or legal action.

The literature on crisis bargaining between rational agents may also help us to understand the circumstances under which compellent threats are made and carried out, and point to mechanisms for avoiding these scenarios. Countering the hypothesis that war between rational agents A and B can occur as a result of indivisible stakes (for example a territory), Powell (2006, p. 178) presents a case similar to that in Example 1.1.1, which shows that allocating the full stakes to each agent according to their probabilities of winning a war Pareto-dominates fighting.

Example 1.1.1 (Simulated conflict).

Consider two countries disputing a territory which has value d for each of them. Suppose that the row country has probability p of winning a conflict, and conflict costs $c > 0$ for each country, so that their payoffs for Surrendering and Fighting are as in the top matrix in Table 2. However, suppose the countries agree on the probability p that the row players win; perhaps they have access to a mutually trusted war-simulator which has row player winning in $100p\%$ of simulations. Then, instead of engaging in real conflict, they could allocate the territory based on a draw from the simulator. Playing this game is preferable, as it saves each country the cost c of actual conflict.

	Surrender	Fight
Surrender	$0, 0$	$0, d$
Fight	$d, 0$	$pd - c, (1-p)d - c$

Conflict

	Surrender	Simulated fight
Surrender	$0, 0$	$0, d$
Simulated fight	$d, 0$	$pd, (1-p)d$

Simulated conflict

Table 2: Allocating indivisible stakes with conflict (top) and simulated conflict (bottom).

If players could commit to the terms of peaceful settlements and truthfully disclose private information necessary for the construction of a settlement (for instance, information pertaining to the outcome probability p in Example 1.1.1), the allocation of indivisible stakes could often be accomplished. Thus, the most plausible of Fearon's rationalist explanations for war seem to be (1) the difficulty of credible commitment and (2) incomplete information (and incentives to misrepresent that information). Section 3° concerns discussion of credibility in TAI systems. In Section 4° we discuss several issues related to the resolution of conflict under private information.

Lastly, while game theory provides a powerful framework for modeling cooperation failure, TAI systems or their operators will not necessarily be well-modeled as rational agents. For example, systems involving humans in the loop, or black-box TAI agents trained by evolutionary methods, may be governed by a complex network of decision-making heuristics not easily captured in a utility function. We discuss research directions that are particularly relevant to cooperation failures among these kinds of agents in Sections 5.2 (Multi-agent training) and 6 (Humans in the loop).

1.2 Outline of the agenda

We list the sections of the agenda below. Different sections may appeal to readers from different backgrounds. For instance, Section 5° (Contemporary AI architectures) may be most interesting to those with some interest in machine learning, whereas Section 7 (Foundations of rational agency) ° will be more relevant to readers with an interest in formal epistemology or the philosophical foundations of decision theory. Tags after the description of

each section indicate the fields most relevant to that section. Some sections contain Examples illustrating technical points, or explaining in greater detail a possible research direction.

- **Section 2: AI strategy and governance.** The nature of losses from cooperation failures will depend on the strategic landscape at the time TAI is deployed. This includes, for instance, the extent to which the landscape is uni- or multipolar (Bostrom, 2014) and the balance between offensive and defensive capabilities (Garfinkel and Dafoe, 2019). Like others with an interest in shaping TAI for the better, we want to understand this landscape, especially insofar as it can help us to identify levers for preventing catastrophic cooperation failures. Given that much of our agenda consists of theoretical research, an important question for us to answer is whether and how such research translates into the governance of TAI.

  Public policy; International relations; Game theory; Artificial intelligence

- **Section 3: Credibility.** Credibility --- for instance, the credibility of commitments to honor the terms of settlements, or to carry out threats --- is a crucial feature of strategic interaction. Changes in agents' ability to self-modify (or create successor agents) and to verify aspects of one another's internal workings are likely to change the nature of credible commitments. These anticipated developments call for the application of existing decision and game theory to new kinds of agents, and the development of new theory (such as that of program equilibrium (Tennenholtz, 2004)) that better accounts for relevant features of machine intelligence.

  Game theory; Behavioral economics; Artificial intelligence

- **Section 4: Peaceful bargaining mechanisms.** Call a *peaceful bargaining mechanism* a set of strategies for each player that does not lead to destructive conflict, and which each agent prefers to playing a strategy which does lead to destructive conflict. In this section, we discuss several possible such strategies and problems which need to be addressed in order to ensure that they are implemented. These strategies include bargaining strategies taken from or inspired by the existing literature on rational crisis bargaining (see Section 1.1, as well as a little-discussed proposal for deflecting compellent threats which we call *surrogate goals* (Baumann, 2017, 2018).

  Game theory; International relations; Artificial intelligence

- **Section 5: Contemporary AI architectures.** Multi-agent artificial intelligence is not a new field of study, and cooperation is of increasing interest to machine learning researchers (Leibo et al., 2017; Foerster et al., 2018; Lerer and Peysakhovich, 2017; Hughes et al., 2018; Wang et al., 2018). But there remain unexplored avenues for

understanding cooperation failures using existing tools for artificial intelligence and machine learning. These include the implementation of approaches to improving cooperation which make better use of agents' potential transparency to one another; the implications of various multi-agent training regimes for the behavior of AI systems in multi-agent settings; and analysis of the decision-making procedures implicitly implemented by various reinforcement learning algorithms.

Machine learning; Game theory

- **Section 6: Humans in the loop°.** Several TAI scenarios and proposals involve a human in the loop, either in the form of a human-controlled AI tool, or an AI agent which seeks to adhere to the preferences of human overseers. These include Christiano (2018c)'s iterated distillation and amplification (IDA; see Cotra 2018 for an accessible introduction), Drexler (2019)'s comprehensive AI ser-vices, and the reward modeling approach of Leike et al. (2018). We would like a better understanding of behavioral game theory, targeted at improving cooperation in TAI landscapes involving human-in-the-loop systems. We are particularly interested in advancing the study of the behavioral game theory of interactions between humans and AIs.

  Machine learning; Behavioral economics

- **Section 7: Foundations of rational agency°.** The prospect of TAI foregrounds several unresolved issues in the foundations of rational agency. While the list of open problems in decision theory, game theory, formal epistemology, and the foundations of artificial intelligence is long, our focus includes decision theory for computationally bounded agents; and prospects for the rationality and feasibility of various kinds of decision-making in which agents take into account non-causal dependences between their actions and their outcomes.

  Formal epistemology; Philosophical decision theory; Artificial intelligence

2 AI strategy and governance [2]

We would like to better understand the ways the strategic landscape among key actors (states, AI labs, and other non-state actors) might look at the time TAI systems are deployed, and to identify levers for shifting this landscape towards widely beneficial outcomes. Our interests here overlap with Dafoe (2018)'s AI governance research agenda (see especially the "Technical Landscape" section), though we are most concerned with questions relevant to risks associated with cooperation failures.

2.1 Polarity and transition scenarios

From the perspective of reducing risks from cooperation failures, it is *prima facie* preferable if the transition to TAI results in a unipolar rather than a distributed outcome: The greater the chances of a single dominant actor, the lower the chances of conflict (at least after that actor has achieved dominance). But the analysis is likely not so simple, if the international relations literature on the relative safety of different power distributions (e.g., Deutsch and Singer 1964; Waltz 1964; Christensen and Snyder 1990) is any indication. We are therefore especially interested in a more fine-grained analysis of possible developments in the balance of power. In particular, we would like to understand the likelihood of the various scenarios, their relative safety with respect to catastropic risk, and the tractability of policy interventions to steer towards safer distributions of TAI-related power. Relevant questions include:

- One might expect rapid jumps in AI capabilities, rather than gradual progress, to make unipolar outcomes more likely. Should we expect rapid jumps in capabilities or are the capability gains likely to remain gradual (AI Impacts, 2018)?
- Which distributions of power are, all things considered, least at risk of catastrophic failures of cooperation?
- Suppose we had good reason to believe we ought to promote more uni- (or multi-) polar outcomes. What are the best policy levers for increasing the concentration (or spread) of AI capabilities, without severe downsides (such as contributing to arms-race dynamics)?

2.2 Commitment and transparency [3][4]

Agents' ability to make credible commitments is a critical aspect of multi-agent systems. Section 3° is dedicated to technical questions around credibility, but it is also important to consider the strategic implications of credibility and commitment.

One concerning dynamic which may arise between TAI systems is *commitment races* (Kokotajlo, 2019a). In the game of Chicken (Table 1), both players have reason to commit to driving ahead as soon as possible, by conspicuously throwing out their steering wheels. Likewise, AI agents (or their human overseers) may want to make certain commitments (for instance, commitments to carry through with a threat if their demands aren't met) as soon as possible, in order to improve their bargaining positions. As with Chicken, this is a dangerous situation. Thus we would like to explore possibilities for curtailing such dynamics.

- At least in some cases, greater transparency seems to limit possibilities for agents to make dangerous simultaneous commitments. For instance, if one country is carefully monitoring another, they are likely to detect efforts to build doomsday devices with which they can make credible commitments. On the other hand, transparency seems to promote the ability to make dangerous commitments: I have less reason to throw out my steering wheel if you can't see me do it. Under what circumstances does mutual transparency mitigate or exacerbate commitment race dynamics, and how can this be used to design safer AI governance regimes?
- What policies can make the success of greater transparency between TAI systems more likely (to the extent that this is desirable)? Are there path dependencies which must be addressed early on in the engineering of TAI systems so that open-source interactions are feasible?

Finally, in human societies, improvements in the ability to make credible commitments (e.g., to sign contracts enforceable by law) seem to have facilitated large gains from trade through more effective coordination, longer-term cooperation, and various other mechanisms (e.g., Knack and Keefer 1995; North 1991; Greif et al. 1994; Dixit 2003).

- Which features of increased credibility promote good outcomes? For instance, laws typically don't allow a threatener to publicly request they be locked up if they don't carry out their threat. How much would societal outcomes change given indiscriminate ability to make credible commitments? Have there been situations where laws and norms around what one can commit to were different from what we see now, and what were the consequences?
- How have past technological advancements changed bargaining between human actors? (Nuclear weapons are one obvious example of a technological advancement which considerably changed the bargaining dynamics between powerful actors.)
- Open-source game theory, described in Section 3.2°, is concerned with an idealized form of mutual auditing. What do historical cases tell us about the factors for the success of mutual auditing schemes? For instance, the Treaty on Open Skies, in which members states agreed to allow unmanned overflights in order to monitor their military activities (Britting and Spitzer, 2002),, is a notable example of such a scheme. See also the literature on "confidence-building" measures in international security, e.g., Landau and Landau (1997) and references therein.
- What are the main costs from increased commitment ability?

2.3 AI misalignment scenarios

Christiano (2018a) defines "the alignment problem" as "the problem of building powerful AI systems that are aligned with their operators". Related problems, as discussed by Bostrom (2014), include the "value loading" (or "value alignment") problem (the problem of ensuring that AI systems have goals compatible with the goals of humans), and the "control problem" (the general problem of controlling a powerful AI agent). Despite the recent surge in attention on AI risk, there are few detailed descriptions of what a future with misaligned AI systems might look like (but see Sotala 2018; Chris-tiano 2019; Dai 2019 for examples). Better models of the ways in which misaligned AI systems could arise and how they might behave are important for our understanding of critical interactions among powerful actors in the future.

- Is AI misalignment more likely to constitute a "near-miss" with respect to human values, or extreme departures from human goals (cf. Bostrom (2003)'s "paperclip maximizer")?
- Should we expect human-aligned AI systems be able to cooperate with misaligned systems (cf. Shulman (2010))?
- What is the likelihood that outright-misaligned AI agents will be deployed alongside aligned systems, versus the likelihood that aligned systems eventually become misaligned by failing to preserve their original goals? (cf. discussion of "goal preservation" (Omohundro, 2008)).
- What does the landscape of possible cooperation failures look like in each of the above scenarios?

2.4 Other directions

According to the *offense-defense theory*, the likelihood and nature of conflict depend on the relative efficacy of offensive and defensive security strategies (Jervis, 2017, 1978; Glaser, 1997). Technological progress seems to have been a critical driver of shifts in the offense-defense balance (Garfinkel and Dafoe, 2019), and the advent of powerful AI systems in strategic domains like computer security or military technology could lead to shifts in that balance.

- To better understand the strategy landscape at the time of AI deployment, we would like to be able to predict technology-induced changes in the offense-defense balance and how they might affect the nature of conflict. One area of interest, for instance, is cybersecurity (e.g., whether leading developers of TAI systems would be able to protect against cyberattacks; cf. Zabel and Muehlhauser 2019).

Besides forecasting future dynamics, we are curious as to what lessons can be drawn from case studies of cooperation failures, and policies which have mitigated or exacerbated such risks. For example: Cooperation failures among powerful agents representing human values may be particularly costly when threats are involved. Examples of possible case studies include nuclear deterrence, ransomware (Gazet, 2010) and its implications for computer security, the economics of hostage-taking (Atkin-son et al., 1987; Shortland and Roberts, 2019), and extortion rackets (Superti, 2009). Such case studies might investigate costs to the threateners, gains for the threateners, damages to third parties, factors that make agents more or less vulnerable to threats, existing efforts to combat extortionists, etc. While it is unclear how informative such case studies will be about interactions between TAI systems, they may be particularly relevant in humans-in-the-loop scenarios (Section 6°).

Lastly, in addition to case studies of cooperation failures themselves, it would be helpful for the prioritization of the research directions presented in this agenda to study how other instances of formal research have influenced (or failed to influence) critical real-world decisions. Particularly relevant examples include the application of game theory to geopolitics (see Weintraub (2017) for a review of game theory and decision-making in the Cold War); cryptography to computer security, and formal verification in the verification of software programs.

2.5 Potential downsides of research on cooperation failures

The remainder of this agenda largely concerns technical questions related to interactions involving TAI-enabled systems. A key strategic question running throughout is: What are the potential downsides to increased technical understanding in these areas? It is possible, for instance, that technical and strategic insights related to credible commitment increase rather than decrease the efficacy and likelihood of compellent threats. Moreover, the naive application of idealized models of rationality may do more harm than good; it has been argued that this was the case in some applications of formal methods to Cold War strategy, for instance Kaplan (1991). Thus the exploration of the dangers and limitations of technical and strategic progress is itself a critical research direction.

The next post in the sequence, "Sections 3 & 4: Credibility, Peaceful Bargaining Mechanisms", will come out Tuesday, December 17.

Acknowledgements & References°

1. The security dilemma refers to a situation in which actions taken by one state to improve their security (e.g., increasing their military capabilities) leads other states to act similarly. This leads to an increase in tensions which all parties would prefer to avoid. ↩
2. Notes by Lukas Gloor contributed substantially to the content of this section. ↩
3. We refer the reader to Garfinkel (2018)'s review of recent developments in cryptography and their possible long-term consequences. The sections of Garfinkel (2018) particularly relevant to issues concerning the transparency of TAI systems and implications for cooperation are sections 3.3 (non-intrusive agreement verification), 3.5 (collective action problems), 4 (limitations and skeptical views on implications of cryptographic technology), and the appendix (relevance of progress in artificial intelligence). See also Kroll et al. (2016)'s review of potential applications of computer science tools, including software verification, cryptopraphic commitments, and zero knowledge proofs, to the accountability of algorithmic decisions. Regarding the problem of ensuring that automated decision systems are "accountable and governable", they write: "We challenge the dominant position in the legal literature that transparency will solve these problems. Disclosure of source code is often neither necessary (because of alternative techniques from computer science) nor sufficient (because of the issues of analyzing code) to demonstrate the fairness of a process." ↩
4. Parts of this subsection were developed from notes by Anni Leskelä. ↩

Center on Long-Term Risk (CLR) 5 | Risks of Astronomical Suffering (S-risks) 3 | Research Agendas 2 | Game Theory 2 | Coordination / Cooperation 2 | AI 2 | Frontpage

Previous:

Preface to CLR's Research Agenda on Cooperation, Conflict, and TAI

7 comments 62 karma

Next:

Sections 3 & 4: Credibility, Peaceful Bargaining Mechanisms

No comments 20 karma

Log in to save where you left off

Mentioned in

53 [Intro to brain-like-AGI safety] 1. What's the problem & Why work on it now?
33 AGI Safety Fundamentals curriculum and application
19 Preface to CLR's Research Agenda on Cooperation, Conflict, and TAI
8 Sections 5 & 6: Contemporary Architectures, Humans in the Loop

6 Sections 3 & 4: Credibility, Peaceful Bargaining Mechanisms

Load More (5/6)

2 comments, sorted by top scoring

[−] **IndraG** 9mo 0

> threats being harmful to their target, the execution of threats seems to constitute an inefficiency:

also when the threats are targeted at (Pareto-dominated) inefficiency, i.e. at (conditional on) any actions other than (coordinating on) the most efficient?

As a concrete counter-example, there are productivity/self-control tools, wherewith people elect to target and/or execute threats on themselves to help elicit better behavior. The legal system is basically also collective threats that help us behave better, but is it inefficient such that we should do better without? I think the opposite, such that any one threat can not only be harmful, but also neutral or beneficial.

[−] **IndraG** 9mo 0

> and a defecting equilibrium which does not.

Why doesn't this also require coordination? Also, there also seems to be a mixed equilibrium where both players randomize their strategies 50/50.

For the Chicken game, the mixed strategy equilibrium is not 50/50 but more specifically 90/10. And a mutual defection can also not constitute crash, but instead allow for further and stable repeated play, particularly if a mixed strategy equilibrium is coordinated and acted upon.

Moderation Log