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

Jesse Clifton
Center on Long-Term Risk

March 4, 2020

# Contents

# 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.

| | | 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).

## 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 $\text{fear} > 0$.

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; Fey and 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. In particular, consider cases where the agents anticipate a swing in the balance of power. In some such cases, the agents who will lose power cannot credibly commit not to take preemptive action, because both parties foresee that they prefer fighting to the default course where they lose power;
- 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 1.0.1, which shows that allocating the full stakes to each agent according to their probabilities of winning a war Pareto-dominates fighting.

[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.

| | 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).

**Example 1.0.1** (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.

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 1.0.1), 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 incomplete 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 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 services, 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]

[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.

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 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; Christiano 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 (Atkinson 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.

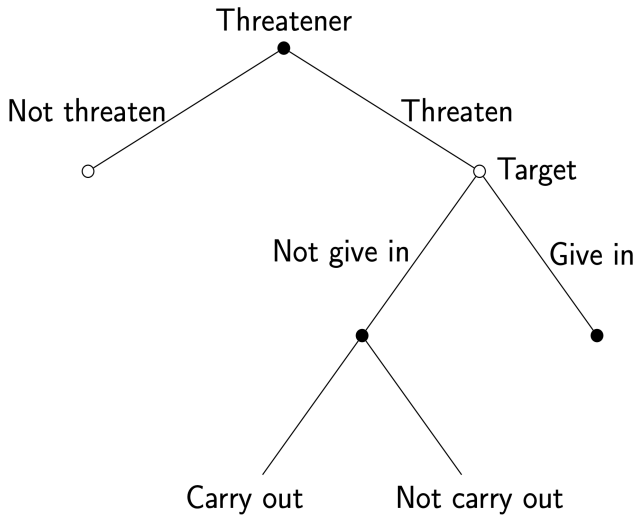


Figure 1: At time 1, Threatener decides to make a threat or not. If a threat is made, Target at time 2 decides to give in or not. If they don't give in, Threatener decides to carry out the threat or not, determining the players' payoffs. Target may reason that Threatener will not carry out the threat if Target doesn't give in, because it is costly for Threatener to do so and cannot affect Target's choice. Therefore, Target won't give in. This is the informal reasoning behind SPE.

# 3 Credibility

*Credibility* is a central issue in strategic interaction. By credibility, we refer to the issue of whether one agent has reason to believe that another will do what they say they will do. Credibility (or lack thereof) plays a crucial role in the efficacy of contracts (Fehr et al., 1997; Bohnet et al., 2001), negotiated settlements for avoiding destructive conflict (Powell, 2006), and commitments to carry out (or refuse to give in to) threats (e.g., Kilgour and Zagare 1991; Konrad and Skaperdas 1997).

In game theory, the fact that Nash equilibria (Section 1.1) sometimes involve *non-credible threats* motivates a refined solution concept called *subgame perfect equilibrium (SPE)*. An SPE is a Nash equilibrium of an extensive-form game in which a Nash equilibrium is also played at each subgame. In the threat game depicted in Figure 1, "carry out" is not played in a SPE, because the threatener has no reason to carry out the threat once the threatened party has refused to give in; that is, "carry out" isn't a Nash equilibrium of the subgame played after the threatened party refuses to give in. So in an SPE-based analysis of one-shot threat situations between rational agents, threats are never carried out because they are not credible (i.e., they violate subgame perfection).

However, agents may establish credibility in the case of repeated interactions by repeatedly making good on their claims (Sobel, 1985). Secondly, despite the fact that carrying out a threat in the one-shot threat game violates subgame perfection, it is a well-known result from behavioral game theory that humans typically refuse unfair

splits in the Ultimatum Game [4] (Güth et al., 1982; Henrich et al., 2006), which is equivalent to carrying out the threat in the one-shot threat game. So executing commitments which are irrational (by the SPE criterion) may still be a feature of human-in-the-loop systems (Section 6), or perhaps systems which have some humanlike game-theoretic heuristics in virtue of being trained in multi-agent environments (Section 5.2). Lastly, threats may become credible if the threatener has *credibly committed* to carrying out the threat (in the case of the game in 1, this means convincing the opponent that they have removed the option (or made it costly) to "Not carry out"). There is a considerable game-theoretic literature on credible commitment, both on how credibility can be achieved (Schelling, 1960) and on the analysis of games under the assumption that credible commitment is possible (Von Stackelberg, 2010; Nash, 1953; Muthoo, 1996; Bagwell, 1995).

## 3.1 Commitment capabilities

It is possible that TAI systems may be relatively transparent to one another; capable of self-modifying or constructing sophisticated commitment devices; and making various other "computer-mediated contracts" (Varian, 2010); see also the lengthy discussions in Garfinkel and Dafoe (2019) and Kroll et al. (2016), discussed in Footnote 1, of potential implications of cryptographic technology for credibility. We want to understand how plausible changes in the ability to make credible commitments affect risks from cooperation failures.

- In what ways does artificial intelligence make credibility more difficult, rather than less so? For instance, AIs lack evolutionarily established mechanisms (like credible signs of anger; Hirshleifer 1987) for signaling their intentions to other agents.
- The credibility of an agent's stated commitments likely depends on how interpretable[5] that agent is to others. What are the possible ways in which interpretability may develop, and how does this affect the propensity to make commitments? For instance, in trajectories where AI agents are increasingly opaque to their overseers, will these agents be motivated to make commitments while they are still interpretable enough to overseers that these commitments are credible?
- In the case of training regimes involving the imitation of human exemplars (see Section 6), can humans also make credible commitments on behalf of the AI system which is imitating them?

## 3.2 Open-source game theory

Tennenholtz (2004) introduced *program games*, in which players submit programs that

[4] The Ultimatum Game is the 2-player game in which Player 1 proposes a split $(pX, (1-p)X)$ of an amount of money $X$, and Player 2 accepts or rejects the split. If they accept, both players get the proposed amount, whereas if they reject, neither player gets anything. The unique SPE of this game is for Player 1 to propose as little as possible, and for Player 2 to accept the offer.

[5] See Lipton (2016); Doshi-Velez and Kim (2017) for recent discussions of interpretability in machine learning.

have access to the source codes of their counterparts. Program games provide a model of interaction under mutual transparency. Tennenholtz showed that in the Prisoner's Dilemma, both players submitting Algorithm 1 is a *program equilibrium* (that is, a Nash equilibrium of the corresponding program game). Thus agents may have incentive to participate in program games, as these promote more cooperative outcomes than the corresponding non-program games. For these reasons, program games may be helpful

**Algorithm 1:** Tennenholtz (2004)'s construction of a program equilibrium of the one-shot Prisoner's Dilemma. The program cooperates if its counterpart's program's source code is identical to its own (and thus both players cooperate), and defects otherwise.

**Input:** Program source codes $s_1, s_2$
**if** $s_1 = s_2$ **then**
  **return** Cooperate
**end**
**else**
  **return** Defect
**end**

to our understanding of interactions among advanced AIs.

Other models of strategic interaction between agents who are transparent to one another have been studied (more on this in Section 5); following Critch (2019), we will call this broader area *open-source game theory*. Game theory with source-code transparency has been studied by Fortnow 2009; Halpern and Pass 2018; LaVictoire et al. 2014; Critch 2019; Oesterheld 2019, and models of multi-agent learning under transparency are given by Brafman and Tennenholtz (2003); Foerster et al. (2018). But open-source game theory is in its infancy and many challenges remain[6].

- The study of program games has, for the most part, focused on the simple setting of two-player, one-shot games. How can (cooperative) program equilibrium strategies be automatically constructed in general settings?
- Under what circumstances would agents be incentivized to enter into open-source interactions?
- How can program equilibrium be made to promote more efficient outcomes even in cases of incomplete access to counterparts' source codes?
  - As a toy example, consider two robots playing a single-shot program prisoner's dilemma, in which their respective moves are indicated by a simultaneous button press. In the absence of verification that the output of the source code actually causes the agent to press the button, it is possible that the output of the program does not match the actual physical action taken. What are the prospects for closing such "credibility gaps"? The literature

[6] See also Section 5.1 for discussion of open-source game theory in the context of contemporary machine learning, and Section 2 for policy considerations surrounding the implementation of open-source interaction.

on (physical) zero-knowledge proofs (Fisch et al., 2014; Glaser et al., 2014) may be helpful here.

– See also the discussion in Section 5.1 on multi-agent learning under varying degrees of transparency.

# 4 Peaceful bargaining mechanisms

In other sections of the agenda, we have proposed research directions for improving our general understanding of cooperation and conflict among TAI systems. In this section, on the other hand, we consider several families of strategies designed to actually avoid catastrophic cooperation failure. The idea of such "peaceful bargaining mechanisms" is, roughly speaking, to find strategies which are 1) peaceful (i.e., avoid conflict) and 2) preferred by rational agents to non-peaceful strategies[7].

We are not confident that peaceful bargaining mechanisms will be used by default. First, in human-in-the-loop scenarios, the bargaining behavior of TAI systems may be dictated by human overseers, who we do not expect to systematically use rational bargaining strategies (Section 6.1). Even in systems whose decision-making is more independent of humans', evolution-like training methods could give rise to non-rational human-like bargaining heuristics (Section 5.2). Even among rational agents, because there may be many cooperative equilibria, additional mechanisms for ensuring coordination may be necessary to avoid conflict arising from the selection of different equilibria (see Example 4.1.1). Finally, the examples in this section suggest that there may be path-dependencies in the engineering of TAI systems (for instance, in making certain aspects of TAI systems more transparent to their counterparts) which determine the extent to which peaceful bargaining mechanisms are available.

In the first subsection, we present some directions for identifying mechanisms which could implement peaceful settlements, drawing largely on existing ideas in the literatures on rational bargaining. In the second subsection we sketch a proposal for how agents might mitigate downsides from threats by effectively modifying their utility function. This proposal is called *surrogate goals*.

## 4.1 Rational crisis bargaining

As discussed in Section 1.1, there are two standard explanations for war among rational agents: credibility (the agents cannot credibly commit to the terms of a peaceful settlement) and incomplete information (the agents have differing private information which makes each of them optimistic about their prospects of winning, and incentives not to disclose or to misrepresent this information).

Fey and Ramsay (2011) model crisis bargaining under incomplete information. They show that in 2-player crisis bargaining games with voluntary agreements (players are able to reject a proposed settlement if they think they will be better off going to war); mutually known costs of war; unknown types $\theta_1, \theta_2$ measuring the players' military strength; a commonly known function $p(\theta_1, \theta_2)$ giving the probability of player 1 winning when the true types are $\theta_1, \theta_2$; and a common prior over types; a peaceful settlement exists if and only if the costs of war are sufficiently large. Such a settlement must compensate each player's strongest possible type by the amount they expect to gain in war.

Potential problems facing the resolution of conflict in such cases include:

[7] More precisely, we borrow the term "peaceful bargaining mechanisms" from Fey and Ramsay (2011), for whom a "peaceful mechanism" is a mapping from each player's type to a payoff such that the probability of war is 0 for every set of types

- Reliance on common prior $\mu$ and agreed-upon win probability model $p(\theta_1, \theta_2)$. If players disagree on these quantities it is not clear how bargaining will proceed. How can players come to an agreement on these quantities, without generating a regress of bargaining problems? One possibility is to defer to a mutually trusted party to estimate these quantities from publicly observed data. This raises its own questions. For example, what conditions must a third party satisfy so that their judgements are trusted by each player? (Cf. Kydd (2003), Rauchhaus (2006), and sources therein on mediation.)

- The exact costs of conflict to each player $c_i$ are likely to be private information, as well. The assumption of a common prior, or the ability to agree upon a prior, may be particularly unrealistic in the case of costs.

Recall that another form of cooperation failure is the simultaneous commitment to strategies which lead to catastrophic threats being carried out (Section 2.2). Such "commitment games" may be modeled as a game of Chicken (Table 1), where Defection corresponds to making commitments to carry out a threat if one's demands are not met, while Cooperation corresponds to not making such commitments. Thus we are interested in bargaining strategies which avoid mutual Defection in commitment games. Such a strategy is sketched in Example 4.0.1.

**Example 4.0.1** (Careful commitments)**.** Consider two agents with access to commitment devices. Each may decide to commit to carrying out a threat if their counterpart does not forfeit some prize (of value 1 to each party). As before, call this decision $D$. However, they may instead commit to carrying out their threat only if their counterpart does not agree to a certain *split* of the prize (say, a split in which Player 1 gets $p$). Call this commitment $C_p$, for "cooperating with split $p$".

When would an agent prefer to make the more sophisticated commitment $C_p$? In order to say whether an agent expects to do better by making $C_p$, we need to be able to say how well they expect to do in the "original" commitment game where their choice is between $D$ and $C$. This is not straightforward, as Chicken admits three Nash equilibria. However, it may be reasonable to regard the players' expected values under mixed strategy Nash equilibrium as the values they expect from playing this game. Thus, split $p$ could be chosen such that $p$ and $1 - p$ exceed player 1 and 2's respective expected payoffs under the mixed strategy Nash equilibrium. Many such splits may exist. This calls for the selection among $p$, for which we may turn to a bargaining solution concept such as Nash (Nash, 1950) or Kalai-Smorokindsky (Kalai et al., 1975). If each player uses the same bargaining solution, then each will prefer to committing to honoring the resulting split of the prize to playing the original threat game, and carried-out threats will be avoided.

Of course, this mechanism is brittle in that it relies on a single take-it-or-leave-it proposal which will fail if the agents use different bargaining solutions, or have slightly different estimates of each players' payoffs. However, this could be generalized to a commitment to a more complex and robust bargaining procedure, such as an alternating-offers procedure (Rubinstein 1982; Binmore et al. 1986; see Muthoo (1996) for a thorough review of such models) or the sequential cooperative bargaining procedure of Van Damme (1986).

Finally, note that in the case where there is uncertainty over whether each player has a commitment device, sufficiently high stakes will mean that players with commitment devices will still have Chicken-like payoffs. So this model can be straightforwardly extended to cases of where the credibility of a threat comes in degrees. An example of a simple bargaining procedure to commit to is Bayesian version of the Nash bargaining solution (Harsanyi and Selten, 1972).

Lastly, see Kydd (2010)'s review of potential applications of the literature rational crisis bargaining to resolving real-world conflict.

## 4.2 Surrogate goals[8]

In this section we introduce *surrogate goals*, a recent[9] proposal for limiting the downsides from cooperation failures (Baumann, 2017, 2018)[10] . We will focus on the phenomenon of coercive threats (for game-theoretic discussion see Ellsberg (1968); Harrenstein et al. (2007)), though the technique is more general. The proposal is: In order to deflect threats against the things it terminally values, an agent adopts a new (surrogate) goal[11]. This goal may still be threatened, but threats carried out against this goal are benign. Furthermore, the surrogate goal is chosen such that it incentives at most marginally more threats.

In Example 4.0.2, we give an example of an operationalization of surrogate goals in a threat game.

**Example 4.0.2** (Surrogate goals via representatives)**.** Consider the game between Threatener and Target, where Threatener makes a demand of Target, such as giving up some resource. Threatener can — at some cost — commit to carrying out a threat against Target . Target can likewise commit to give in to such threats or not. A simple model of this game is given in the payoff matrix in Table 3 (a normal-form variant of the threat game discussed in Section 3[12]).

Unfortunately, players may sometimes play (Threaten, Not give in). For example, this may be due to uncoordinated selection among the two pure-strategy Nash equilibria ((Give in, Threaten) and (Not give in, Not threaten)).

But suppose that, in the above scenario, Target is capable of certain kinds of credible commitments, or otherwise is represented by an agent, Target's Representative, who is. Then Target or Target's Representative may modify its goal architecture to adopt a *surrogate goal* whose fulfillment is not actually valuable to that player, and which is

[8] This subsection is based on notes by Caspar Oesterheld.

[9] Although, the idea of modifying preferences in order to get better outcomes for each player was discussed by Raub (1990) under the name "preference adaptation", who applied it to the promotion of cooperation in the one-shot Prisoner's Dilemma.

[10] See also the discussion of surrogate goals and related mechanisms in Christiano and Wiblin (2019).

[11] Modifications of an agent's utility function have been discussed in other contexts. Omohundro (2008) argues that "AIs will try to preserve their utility functions" and "AIs will try to prevent counterfeit utility". Everitt et al. (2016) present a formal model of a reinforcement learning agent who is able to modify its utility function, and study conditions under which agents self-modify.

[12] Note that the normal form representation in Table 3 is over-simplifying; it assumes the credibility of threats, which we saw in Section 3 to be problematic. For simplicity of exposition, we will nevertheless focus on this normal-form game in this section.

| | | Threatener | |
|---|---|---|---|
| | | Threaten true goal | Not threaten |
| Target | Give in | $-5, 5$ | $0, 0$ |
| | Not give in | $-10, -2$ | $0, 0$ |

Table 3: Payoff matrix for original threat game in normal form. Target payoffs are in blue for easy comparison with the surrogate game (Table 4).

| | | Threatener | | |
|---|---|---|---|---|
| | | Threaten true goal | Threaten surrogate goal | Not threaten |
| Target | Give in | $-5, -5, 5$ | $-5, -5, 5$ | $0, 0, 0$ |
| | Not give in | $-10, -10, -2$ | $0, -10, -1.9$ | $0, 0, 0$ |

Table 4: Payoff matrix for threat game with surrogate goals in normal form. Payoffs for Target are in blue, while Target representative and Threatener payoffs are in black.

slightly cheaper for Threatener to threaten. (More generally, Target could modify itself to commit to acting as if it had a surrogate goal in threat situations.) If this modification is credible, then it is rational for Threatener to threaten the surrogate goal, obviating the risk of threats against Target's true goals being carried out.

As a first pass at a formal analysis: Adopting an additional threatenable goal adds a column to the payoff matrix, as in Table 4. And this column weakly dominates the old threat column (i.e., the threat against Target's true goals). So a rational player would never threaten Target's true goal. Target does not themselves care about the new type of threats being carried out, so for her, the utilities are given by the blue numbers in Table 4.

This application of surrogate goals, in which a threat game is already underway but players have the opportunity to self-modify or create representatives with surrogate goals, is only one possibility. Another is to consider the adoption of a surrogate goal as the choice of an agent (before it encounters any threat) to commit to acting according to a new utility function, rather than the one which represents their true goals. This could be modeled, for instance, as an extensive-form game of incomplete information in which the agent decides which utility function to commit to by reasoning about (among other things) what sorts of threats having the utility function might provoke. Such models have a signaling game component, as the player must successfully signal to distrustful counterparts that it will actually act according to the surrogate utility function when threatened. The game-theoretic literature on signaling (Kreps and Sobel, 1994) and the literature on inferring preferences in multi-agent settings (Yu et al., 2019; Lin et al., 2019) may suggest useful models. The implementation of surrogate goals faces a number of obstacles. Some problems and questions include:

- The surrogate goal must be credible, i.e., threateners must believe that the agent will act consistently with the stated surrogate goal. TAI systems are unlikely to have easily-identifiable goals, and so must signal their goals to others through

their actions. This raises questions both of how to signal so that the surrogate goal is at all credible, and how to signal in a way that doesn't interfere too much with the agent's true goals. One possibility in the context of Example 4.0.2 is the use of zero-knowledge proofs (Goldwasser et al., 1989; Goldreich and Oren, 1994) to reveal the Target's surrogate goal (but not how they will actually respond to a threat) to the Threatener.

- How does an agent come to adopt an appropriate surrogate goal, practically speaking? For instance, how can advanced ML agents be trained to reason correctly about the choice of surrogate goal?

- The reasoning which leads to the adoption of a surrogate goal might in fact lead to *iterated* surrogate goals. That is, after having adopted a surrogate goal, Target may adopt a surrogate goal to protect *that* surrogate goal, and so on. Given that Threatener must be incentivized to threaten a newly adopted surrogate goal rather than the previous goal, this may result in Target giving up much more of its resources than it would if only the initial surrogate goal were threatened.

- How do surrogate goals interact with open-source game theory (Sections 3.2 and 5.1)? For instance, do open source interactions automatically lead to the use of surrogate goals in some circumstances?

- In order to deflect threats against the original goal, the adoption of a surrogate goal must lead to a similar distribution of outcomes as the original threat game (modulo the need to be slightly cheaper to threaten). Informally, Target should expect Target's Representative to have the same propensity to give in as Target; how this is made precise depends on the details of the formal surrogate goals model.

A crucial step in the investigation of surrogate goals is the development of appropriate theoretical models. This will help to gain traction on the problems listed above.

# 5 Contemporary AI architectures

Although the architectures of TAI systems will likely be quite different to existing ones, it may still be possible to gain some understanding of cooperation failures among such systems using contemporary tools[13]. First, it is plausible that some aspects of contemporary deep learning methods will persist in TAI systems, making experiments done today directly relevant. Second, even if this is not the case, such research may still help by laying the groundwork for the study of cooperation failures in more advanced systems.

## 5.1 Learning to solve social dilemmas

As mentioned above, some attention has recently been devoted to social dilemmas among deep reinforcement learners (Leibo et al., 2017; Peysakhovich and Lerer, 2017; Lerer and Peysakhovich, 2017; Foerster et al., 2018; Wang et al., 2018). However, a fully general, scalable but theoretically principled approach to achieving cooperation among deep reinforcement learning agents is lacking. In Example 5.0.1 we sketch a general approach to cooperation in general-sum games which subsumes several recent methods, and afterwards list some research questions raised by the framework.

**Example 5.0.1** (Sketch of a framework for cooperation in general-sum games.)**.** The setting is a 2-agent decision process. At each timestep $t$, each agent $i$ receives an observation $o_i^t$; takes an action $a_i^t = \pi_i(o_i^t)$ based on their policy $\pi_i$ (assumed to be deterministic for simplicity); and receives a reward $r_i^t$. Player $i$ expects to get a value of $V_i(\pi_1, \pi_2)$ if the policies $\pi_1, \pi_2$ are deployed. Examples of such environments which are amenable to study with contemporary machine learning tools are the "sequential social dilemmas" introduced by Leibo et al. (2017). These include a game involving potential conflict over scarce resources, as well as a coordination game similar in spirit to Stag Hunt (Table 1).

Suppose that the agents (or their overseers) have the opportunity to choose what policies to deploy by simulating from a model, and to bargain over the choice of policies. The idea is for the parties to arrive at a welfare function $w$ which they agree to jointly maximize; deviations from the policies which maximize the welfare function will be punished if detected. Let $V_i^d$ be a "disagreement point" measuring how well agent $i$ expects to do if they deviate from the welfare-maximizing policy profile. This could be their security value $\max_{\pi_1} \min_{\pi_2} V_i(\pi_1, \pi_2)$, or an estimate of their value when the agents use independent learning algorithms. Finally, define player $i$'s ideal point $V_i^* = \arg\max_{\pi_1,\pi_2} V_i(\pi_1, \pi_2)$. Table 5 displays welfare functions corresponding to several widely-discussed bargaining solutions, adapted to the multi-agent reinforcement learning setting.

Define the cooperative policies as $\pi_1^C, \pi_2^C = \arg\max_{\pi_1,\pi_2} w(\pi_1, \pi_2)$. We need a way of detecting defections so that we can switch from the cooperative policy $\pi_1^C$ to a punishment policy. Call a function that detects defections a "switching rule". To make

[13]Cf. Christiano (2016b)'s discussion of "prosaic" artificial general intelligence, defined as that "which doesn't reveal any fundamentally new ideas about the nature of intelligence or turn up any 'unknown unknowns'."

| **Name of welfare function** $w$ | **Form of** $w(\pi_1, \pi_2)$ |
|---|---|
| Nash (Nash, 1950) | $\left[V_1(\pi_1, \pi_2) - V_1^d\right] \cdot \left[V_2(\pi_1, \pi_2) - V_2^d\right]$ |
| Kalai-Smorodinsky (Kalai et al., 1975) | $V_1(\pi_1, \pi_2)^2 + V_2(\pi_1, \pi_2)^2$ $-\iota\left\{\frac{V_1(\pi_1,\pi_2)-V_1^d}{V_2(\pi_1,\pi_2)-V_2^d} = \frac{V_1^*-V_1^d}{V_2^*-V_2^d}\right\}$ |
| Egalitarian (Kalai, 1977) | $\min\left\{V_1(\pi_1, \pi_2) - V_1^d, V_2(\pi_1, \pi_2) - V_2^d\right\}$ |
| Utilitarian | $V_1(\pi_1, \pi_2) + V_2(\pi_1, \pi_2)$ |

Table 5: Welfare functions corresponding to several widely-discussed bargaining solutions, adapted to the multi-agent RL setting where two agents with value functions $V_1, V_2$ are bargaining over the pair of policies $\pi_1, \pi_2$ to deploy. The function $\iota$ in the definition of the Kalai-Smorodinsky welfare is the $\infty$-0 indicator, used to enforce the constraint in its argument. Note that when the space of feasible payoff profiles is convex, the Nash welfare function uniquely satisfies the properties of (1) Pareto optimality, (2) symmetry, (3) invariance to affine transformations, and (4) independence of irrelevant alternatives. The Nash welfare can also be obtained as the subgame perfect equilibrium of an alternating-offers game as the "patience" of the players goes to infinity (Binmore et al., 1986). On the other hand, Kalai-Smorodinsky uniquely satisfies (1)-(3) plus (5) resource monotonicity, which means that all players are weakly better off when there are more resources to go around. The egalitarian solution satisfies (1), (2), (4), and (5). The utilitarian welfare function is implicitly used in the work of Peysakhovich and Lerer (2017); Lerer and Peysakhovich (2017); Wang et al. (2018) on cooperation in sequential social dilemmas.

the framework general, consider switching rules $\chi$ which return 1 for $\mathrm{Switch}$ and 0 for $\mathrm{Stay}$. Rules $\chi$ depend on the agent's observation history $H_i^t$. The contents of $H_i^t$ will differ based on the degree of observability of the environment, as well as how transparent agents are to each other (cf. Table 6). Example switching rules include:

- $\mathrm{Switch}$ when I see that my counterpart doesn't follow the cooperative policy (cf. Lerer and Peysakhovich 2017): $\chi(H_1^t) = \mathbb{1}\left\{a_2^t \neq \pi_2^C(o_2^t)\right\}$;
- $\mathrm{Switch}$ when my rewards indicate my counterpart is not cooperating (Peysakhovich and Lerer, 2017): $\chi(H_1^t) = \mathbb{1}\left\{\frac{1}{t}\sum_{v=1}^{t} r_1^v < V_1(\pi_1^C, \pi_2^C) - \kappa^t\right\}$, for some $\kappa^t > 0$;
- $\mathrm{Switch}$ when the probability that my counterpart is cooperating, according to my trained defection-detecting model, is low (cf. Wang et al. 2018): $\chi(H_1^t) = \mathbb{1}\left\{P\left(\pi_2^C \mid H_1^t\right) < 1 - \delta^t\right\}$, for some $\delta^t \in (0, 1)$.

Finally, the agents need punishment policies $\pi_i^D$ to switch to in order to disincentivize defections. An extreme case of a punishment policy is the one in which an agent commits to minimizing their counterpart's utility once they have defected: $\pi_1^{D,\mathrm{minimax}} = \arg\min_{\pi_1} \max_2 V_2(\pi_1, \pi_2)$. This is the generalization of the so-called "grim trigger" strategy underlying the classical theory of iterated games (Friedman, 1971; Axelrod, 2000). It can be seen that each player submitting a grim trigger strategy in the above framework constitutes a Nash equilibrium in the case that the counterpart's observations and actions are visible (and therefore defections can be detected with certainty). However, grim trigger is intuitively an extremely dangerous strategy for promoting cooperation, and indeed does poorly in empirical studies of different strategies for the iterated Prisoner's Dilemma (Axelrod and Hamilton, 1981). One possibility is to train more forgiving, tit-for-tat-like punishment policies, and play a mixed strategy when choosing which to deploy in order to reduce exploitability.

Some questions facing a framework for solving social dilemmas among deep reinforcement learners, such as that sketched in Example 5.0.1, include:

- How does the ability of agents to cooperate deteriorate as their ability to observe one another's actions is reduced?
- The methods for promoting cooperation among deep reinforcement learners discussed in Example 5.0.1 assume 1) complete information (agents do not have private information about, say, their utility functions) and 2) only two players. How can cooperation be achieved in cases of incomplete information and in coalitional games?

In addition to the theoretical development of open-source game theory (Section 3.2), interactions between transparent agents can be studied using tools like deep reinforcement learning. Learning equilibrium (Brafman and Tennenholtz, 2003) and learning with opponent-learning awareness (LOLA) (Foerster et al., 2018; Baumann et al., 2018; Letcher et al., 2018) are examples of analyses of learning under transparency.

- "Opponent-aware" methods like Foerster et al. (2018)'s LOLA[14] assume that agents can efficiently verify relevant aspects of one another's internal workings. How can such verification be achieved in practice? How can agents still reap some of the benefits of transparency in the case of incomplete verifiability? Table 6 lists several recent multi-agent learning techniques which assume varying degrees of agent transparency; given the difficulty of achieving total transparency, successful real-world auditing schemes will likely require a blend of such techniques.
- How should agents of asymmetric capabilities conduct open-source interactions? (As a simple example, one might consider interactions between a purely model-free agent and one which has access to an accurate world model.)

| **Multi-agent learning technique** | **Elements which are mutually transparent** |
|---|---|
| Consequentialist conditional cooperation (CCC) (Peysakhovich and Lerer, 2017) | Reward function, partially observed state |
| Wang et al. (2018) | Reward function, fully observed state |
| Approximate Markov tit-for-tat (amTFT) | Reward function, action, fully observed state |
| Learning with opponent-learning awareness (LOLA) (Foerster et al., 2018) | Observation history, policy parameters |

Table 6: Several recent approaches to achieving cooperation in social dilemmas, which assume varying degrees of agent transparency. In Peysakhovich and Lerer (2017)'s consequentialist conditional cooperation (CCC), players learn cooperative policies offline by optimizing the total welfare. During the target task, they only partially observe the game state and see none of their counterpart's actions; thus, they use only their observed rewards to detect whether their counterpart is cooperating or defecting, and switch to their cooperative or defecting policies accordingly. On the other hand, in Lerer and Peysakhovich (2017), a player sees their counterpart's action and switches to the defecting policy if that action is consistent with defection (mimicking the tit-for-tat strategy in the iterated Prisoner's Dilemma (Axelrod and Hamilton, 1981)).

## 5.2 Multi-agent training

Multi-agent training is an emerging paradigm for the training of generally intelligent agents (Lanctot et al., 2017; Rabinowitz et al., 2018; Suarez et al., 2019; Leibo et al., 2019). It is as yet unclear what the consequences of such a learning paradigm are for the prospects for cooperativeness among advanced AI systems.

[14] Although Foerster et al. (2018) develop a version of LOLA with "opponent modeling" where an agent only makes inferences about their counterpart's parameters, rather than actually seeing them. Zhang and Lesser (2010) present a similar method, though unlike LOLA theirs does not attempt to shape the counterpart's update.

- Will multi-agent training result in human-like bargaining behaviors, involving for instance the costly punishment of those perceived to be acting unfairly (Henrich et al., 2006)? What are the implications for the relative ability of, say, classical and behavioral game theory[15] to predict the behavior of TAI-enabled systems? And, critically, what are the implications for these agents' ability to implement peaceful bargaining strategies (Section 4)? See especially the literature on behavioral evidence regarding rational crisis bargaining (Quek, 2017; Renshon et al., 2017). See also Section 6.1.

- One potentially significant disanalogy of multi-agent training with human biological and cultural evolution is the possibility that agents will have (partial) access to one another's internal workings (see Sections 3.2 and 5.1). What can experiments in contemporary ML architectures tell us about the prospects for efficiency gains from open-source multi-agent learning (Section 5.1)?

- How interpretable will agents trained via multi-agent training be? What are the implications for their ability to make credible commitments (Section 3)?

- Adversarial training has been proposed as an approach to limiting risks from advanced AI systems (Christiano, 2018d; Uesato et al., 2018). Are risks associated with cooperation failures (such as the propensity to make destructive threats) likely to be found by default adversarial training procedures, or is there a need for the development of specialized techniques?

## 5.3 Decision theory

Understanding the decision-making procedures implemented by different machine learning algorithms may be critical for assessing how they will behave in high-stakes interactions with humans or other AI agents. One potentially relevant factor is the *decision theory* implicitly implemented by a machine learning agent. We discuss decision theory at greater length in Section 7.2, but briefly: By an agent's decision theory, we roughly mean which dependences the agent accounts for when predicting the outcomes of its actions. While it is standard to consider only the causal effects of one's actions ("causal decision theory" (CDT)), there are reasons to think agents should account for non-causal evidence that their actions provide about the world[16]. And, different ways of computing the expected effects of actions may lead to starkly different behavior in multi-agent settings.

- Oesterheld (2017a) considers a simple agent designed to maximize the approval score given to it by an overseer (i.e., "approval-directed" Christiano 2014). He

[15] Rabin (1993); Fehr and Schmidt (1999); Bolton and Ockenfels (2000) study fairness and trust; Camerer and Hua Ho (1999) develop a large class of models for explaining human learning in games; and Camerer (2008, Ch. 4) reviews the behavioral literature on bargaining, concluding that a satisfactory theory of bargaining would "probably weave together perceptions of equity. . . , stable social preferences for equal payoffs or fair treatment, heuristic computation, and individual differences...". Also see the discussion of behavioral game theory and human evolution by Hagen and Hammerstein (2006) and references therein.

[16] See also Camerer and Hua Ho (1999)'s distinction between "the law of actual effect" and "the law of simulated effect".

shows that the decision theory implicit in the decisions of such an agent is determined by how the agent and overseer compute the expected values of actions. In this vein: What decision-making procedures are implicit in ML agents trained according to different protocols? See for instance Krueger et al. (2019)'s discussion of "hidden incentives for inducing distributional shift" associated with certain population-based training methods (Jaderberg et al., 2017) for reinforcement learning; cf. Everitt et al. (2019) on understanding agent incentives with causal influence diagrams.

- A "model-free" agent is one which implicitly learns the expected values of its actions by observing the streams of rewards that they generate; such agents are the focus of most deep reinforcement learning research. By contrast, a "model-based" agent (Sutton and Barto, 2018, Ch. 8) is one which explicitly models the world and computes the expected values of its actions by simulating their effects on the world using this model. In certain model-based agents, an agent's decision theory can be specified directly by the modeler, rather than arising implicity[17]. Do any decision-theoretic settings specified by the modeler robustly lead to cooperative outcomes across a wide range of multi-agent environments? Or are outcomes highly sensitive to the details of the situation?

[17] For example, (Everitt et al., 2015) develop sequential extensions of the most commonly studied decision theories, causal and evidential decision theory, in a general reinforcement learning framework. One could develop similar extensions for model-based multi-agent frameworks, like Gmytrasiewicz and Doshi (2005)'s interactive partially observable Markov decision processes.

# 6 Humans in the loop[18]

TAI agents may acquire their objectives via interaction with or observation of humans. Relatedly, TAI systems may consist of AI-assisted humans, as in Drexler (2019)'s comprehensive AI services scenario. Relevant AI techniques include:

- Approval-directedness, in which an agent attempts to maximize human-assigned approval scores (Akrour et al., 2011; Christiano, 2014);
- Imitation (Schaal, 1999; Ross et al., 2011; Evans et al., 2018), in which an agent attempts to imitate the behavior of a demonstrator;
- Preference inference (Ng et al., 2000; Hadfield-Menell et al., 2016; Christiano et al., 2017; Leike et al., 2018), in which an agent attempts to learn the reward function implicit in the behavior of a demonstrator and maximize this estimated reward function.

In human-in-the-loop scenarios, human responses will determine the outcomes of opportunities for cooperation and conflict.

## 6.1 Behavioral game theory

Behavioral game theory has often found deviations from theoretical solution concepts among human game-players. For instance, people tend to reject unfair splits in the ultimatum game despite this move being ruled out by subgame perfection (Section 3). In the realm of bargaining, human subjects often reach different bargaining solutions than those standardly argued for in the game theory literature (in particular, the Nash (Nash, 1950) and Kalai-Smorodinsky (Kalai et al., 1975) solutions) (Felsenthal and Diskin, 1982; Schellenberg, 1988). Thus the behavioral game theory of human-AI interaction in critical scenarios may be a vital complement to theoretical analysis when designing human-in-the-loop systems.

- Under what circumstances do humans interacting with an artificial agent become convinced that the agent's commitments are credible (Section 3)? How do humans behave when they believe their AI counterpart's commitments are credible or not? Are the literatures on trust and artificial agents (e.g., Grodzinsky et al. 2011; Coeckelbergh 2012) and automation bias (Mosier et al., 1998; Skitka et al., 1999; Parasuraman and Manzey, 2010) helpful here? (See also Crandall et al. (2018), who develop an algorithm for promoting cooperation between humans and machines.)
- In sequential games with repeated opportunities to commit via a credible commitment device, how quickly do players make such commitments? How do other players react? Given the opportunity to commit to bargaining rather than to simply carry out a threat if their demands aren't met (see Example 4.0.1), what do

[18] Notes by Lukas Gloor contributed substantially to the content of this section.

players do? Cf. experimental evidence regarding commitment and crisis bargaining; e.g., Quek (2017) finds that human subjects go to war much more frequently in a war game when commitments are not enforceable.

- Sensitivity to stakes varies over behavioral decision- and game-theoretic contexts (e.g., Kahneman et al. 1999; Dufwenberg and Gneezy 2000; Schmidt et al. 2001; Andersen et al. 2011). How sensitive to stakes are the behaviors in which we are most interested? (This is relevant as we're particularly concerned with *catastrophic* failures of cooperation.)

- How do humans model the reasoning of intelligent computers, and what are the implications for limiting downsides in interactions involving humans? For instance, in experiments on games, humans tend to model their counterparts as reasoning at a lower depth than they do (Camerer et al., 2004)[19]. But this may not be the case when humans instead face computers they believe to be highly intelligent.

- How might human attitudes towards the credibility of artificial agents change over time — for instance, as a result of increased familiarity with intelligent machines in day-to-day interactions? What are the implications of possible changes in attitudes for behavioral evidence collected now?

- We are also interested in extensions of existing experimental paradigms in behavioral game theory to interactions between humans and AIs, especially research on costly failures such as threats (Bolle et al., 2011; Andrighetto et al., 2015).

## 6.2 AI delegates

In one class of TAI trajectories, humans control powerful AI delegates who act on their behalf (gathering resources, ensuring safety, etc.). One model for powerful AI delegates is Christiano (2016a)'s (recursively titled) "Humans consulting HCH" (HCH). Saunders (2019) explains HCH as follows:

> HCH, introduced in Humans consulting HCH (Christiano, 2016a), is a computational model in which a human answers questions using questions answered by another human, which can call other humans, which can call other humans, and so on. Each step in the process consists of a human taking in a question, optionally asking one or more sub-questions to other humans, and returning an answer based on those subquestions. HCH can

[19] This has been illustrated in the $p$-Beauty Contest Game (BCG). In the BCG, multiple players simultaneously say a number between 0 and 100. The winner is the person whose number is closest to the mean of all the numbers, times a commonly known number $p$ in $(0, 1)$. If there is a tie, the payoff is divided evenly. This game has a single Nash equilibrium: everyone says 0. However, human players typically don't play this way. Instead, experimental evidence suggests that players model others as reasoning fewer steps ahead than they ("If the know I choose X then they will choose Y, so then I will choose Z instead…"), and then choose the best response to these predicted moves (Nagel, 1995).

be used as a model for what Iterated Amplification[20] would be able to do in the limit of infinite compute.

A particularly concerning class of cooperation failures in such scenarios are threats by AIs or AI-assisted humans against one another.

- Threats could target 1) the delegate's objectives (e.g., destroying the system's resources or its ability to keep its overseer alive and comfortable), or 2) the human overseer's terminal values. Threats of the second type might be much worse. It seems important to investigate the incentives for would-be threateners to use one type of threat or the other, in the hopes of steering dynamics towards lower-stakes threats.
- We are also interested in how interactions between humans and AI delegates could be limited so as to minimize threat risks.

Saunders also discusses a hypothetical manual for overseers in the HCH scheme. In this manual, overseers could find advice "on how to corrigibly answer questions by decomposing them into sub-questions." Exploring practical advice that could be included in this manual might be a fruitful exercise for identifying concrete interventions for addressing cooperation failures in HCH and other human-in-the-loop settings. Examples include:

- Instructions related to rational crisis bargaining (Section 4.1);
- Instructions related to the implementation of surrogate goals (Section 4.2).

[20] Iterated (Distillation and) Amplification (IDA) is Christiano (2018b)'s proposal for training aligned AI systems. In brief, it consists of iterating a Distillation step in which the capabilities of a team of AI delegates are distilled into a single agent; and an Amplification step, in which the capabilities of the distilled agent are amplified by copying that agent many times and delegating different tasks to different copies. The hope for IDA as an approach to AI safety is that many slightly less-capable agents will be able to control the more powerful agent produced by the latest Distill step, at each iteration of the process. See Cotra (2018) for an accessible overview of IDA.

# 7 Foundations of rational agency

We think that the effort to ensure cooperative outcomes among TAI systems will likely benefit from thorough conceptual clarity about the nature of rational agency. Certain foundational achievements — probability theory, the theory of computation, algorithmic information theory, decision theory, and game theory to name some of the most profound — have been instrumental in both providing a powerful conceptual apparatus for thinking about rational agency, and the development of concrete tools in artificial intelligence, statistics, cognitive science, and so on. Likewise, there are a number of outstanding foundational questions surrounding the nature of rational agency which we expect to yield additional clarity about interactions between TAI-enabled systems. Broadly, we want to answer:

- What are the implications of *computational boundededness* (Russell and Subramanian, 1994; Cherniak, 1984; Gershman et al., 2015) for normative decision theory, in particular as applied to interactions among TAI systems?
- How should agents handle non-causal dependences with other agents' decision-making in their own decisions?

We acknowledge, however, the limitations of the agenda for foundational questions which we present. First, it is plausible that the formal tools we develop will be of limited use in understanding TAI systems that are actually developed. This may be true of black-box machine learning systems, for instance[21]. Second, there is plenty of potentially relevant foundational inquiry scattered across epistemology, decision theory, game theory, mathematics, philosophy of probability, philosophy of science, etc. which we do not prioritize in our agenda[22]. This does not necessarily reflect a considered judgement about all relevant areas. However, it is plausible to us that the research directions listed here are among the most important, tractable, and neglected (Concepts, n.d.) directions for improving our theoretical picture of TAI.

## 7.1 Bounded decision theory

Bayesianism (Talbott, 2016) is the standard idealized model of reasoning under empirical uncertainty. Bayesian agents maintain probabilities over hypotheses; update these probabilities by conditionalization in light of new evidence; and make decisions according to some version of expected utility decision theory (Briggs, 2019). But Bayesianism faces a number of limitations when applied to computationally bounded agents. Examples include:

[21] Cf. discussion of the Machine Intelligence Research Institute foundational research and its applicability to machine-learning-driven systems Taylor (2016); Dewey (2017).

[22] For other proposals for foundational research motivated by a concern with improving the long-term future, see for instance the research agendas of the Global Priorities Research Institute (Greaves et al., 2019) (especially Sections 2.1 and 2.2 and Appendix B) and the Machine Intelligence Research Institute (Soares and Fallenstein, 2017; Garrabrant and Demski, 2018).

[22] This subsection was developed from an early-stage draft by Caspar Oesterheld and Johannes Treutlein.

- Unlike Bayesian agents, computationally bounded agents are *logically uncertain*. That is, they are not aware of all the logical implications of their hypotheses and evidence (Garber, 1983)[23]. Logical uncertainty may be particularly relevant in developing a satisfactory open-source game theory (Section 3.2), as open-source game theory requires agents to make decisions on the basis of the output of their counterparts' source codes (which are logical facts). In complex settings, agents are unlikely to be certain about the output of all of the relevant programs. Garrabrant et al. (2016) presents a theory for assigning logical credences, but it has flaws when applied to decision-making (Garrabrant, 2017). Thus one research direction we are interested in is a theoretically sound and computationally realistic approach to decision-making under logical uncertainty.

- Unlike Bayesian agents, computationally bounded agents cannot reason over the space of all possible hypotheses. Using the the terminology of statistical modeling (e.g., Hansen et al. 2016), we will call this situation *model misspecification*[24].The development of a decision theory for agents with misspecified world-models would seem particularly important for our understanding of *commitment* in multi-agent settings. Rational agents may sometimes want to bind themselves to certain policies in order to, for example, reduce their vulnerability to exploitation by other agents (e.g., Schelling (1960); Meacham (2010); Kokotajlo (2019a); see also Section 3 and the discussion of commitment races in Section 2). Intuitively, however, a rational agent may be hesitant to bind themselves to a policy by planning with a model which they suspect is misspecified. The analysis of games of incomplete information may also be quite sensitive to model misspecification[25]. To develop a better theory of reasoning under model misspecification, one might start with the literatures on decision theory under ambiguity (Gilboa and Schmeidler, 1989; Maccheroni et al., 2006; Stoye, 2011; Etner et al., 2012) and robust control theory (Hansen and Sargent, 2008).

## 7.2 Acausal reasoning[26]

[23]Consider, for instance, that most of us are uncertain about the value of the $10^{10^{th}}$ digit of $\pi$, despite the fact that its value logically follows from what we know about mathematics.

[24]This problem has been addressed in two ways. The first is simply to posit that the agent reasons over an extremely rich class of hypotheses, perhaps one rich enough to capture all of the important possibilities. An example of such a theory is Solomonoff induction (Solomonoff, 1964; Sterkenburg, 2013), in which evidence takes the form of a data stream received via the agent's sensors, and the hypotheses correspond to all possible "lower semi-computable" generators of such data streams. But Solomonoff induction is incomputable and its computable approximations are still intractable. The other approach is to allow agents to have incomplete sets of hypotheses, and introduce an additional rule by which hypotheses may be added to the hypothesis space (Wenmackers and Romeijn, 2016). This sort of strategy seems to be the way forward for an adequate theory of bounded rationality in the spirit of Bayesianism. However, to our knowledge, there is no *decision theory* which accounts for possible amendments to the agent's hypothesis space.

[25]See Section 4.1 for discussion of games of incomplete information and possible limitations of Bayesian games.

[26]This subsection was developed from an early-stage draft by Daniel Kokotajlo and Johannes Treutlein.

Newcomb's problem[27] (Nozick, 1969) showed that classical decision theory bifurcates into two conflicting principles of choice in cases where outcomes depend on agents' predictions of each other's behavior. Since then, considerable philosophical work has gone towards identifying additional problem cases for decision theory and towards developing new decision theories to address them. As with Newcomb's problem, many decision-theoretic puzzles involve dependences between the choices of several agents. For instance, Lewis (1979) argues that Newcomb's problem is equivalent to a prisoner's dilemma played by agents with highly correlated decision-making procedures, and Soares and Fallenstein (2015) give several examples in which artificial agents implementing certain decision theories are vulnerable to blackmail.

In discussing the decision theory implemented by an agent, we will assume that the agent maximizes some form of expected utility. Following Gibbard and Harper (1978), we write the expected utility given an action $a$ for a single-stage decision problem in context $x$ as

$$EU(a) \triangleq \sum_j P(a \to o_j; x)u(o_j), \tag{1}$$

where $o_j$ are possible outcomes; $u$ is the agent's utility function; and $\to$ stands for a given notion of dependence of outcomes on actions. The dependence concept an agent uses for $\to$ in part determines its decision theory.

The philosophical literature has largely been concerned with *causal decision theory (CDT)* (Gibbard and Harper, 1978) and *evidential decision theory (EDT)* (Horgan, 1981), which are distinguished by their handling of dependence.

Causal conditional expectations account only for the causal effects of an agent's actions; in the formalism of Pearl (2009)'s do-calculus, for instance, the relevant notion of expected utility conditional on action $a$ is $\mathbb{E}[U \mid \mathrm{do}(a)]$. EDT, on the other hand, takes into account non-causal dependencies between the agent's actions and the outcome. In particular, it takes into account the evidence that taking the action provides for the actions taken by *other* agents in the environment with whom the decision-maker's actions are dependent. Thus the evidential expected utility is the classical conditional expectation $\mathbb{E}[U \mid A = a]$.

Finally, researchers in the AI safety community have more recently developed what we will refer to as *logical* decision theories, which employ a third class of dependence for evaluating actions (Dai, 2009; Yudkowsky, 2009; Yudkowsky and Soares, 2017). One such theory is functional decision theory (FDT)[28], which uses what Yudkowsky

[27] In Newcomb's problem, a player is faced with two boxes: a clear box which contains \$1000, and an opaque box which contains either \$0 or \$1 million. They are given a choice between choosing both boxes (Two-Boxing) or choosing only the opaque box (One-Boxing). They are told that, before they were presented with this choice, a highly reliable predictor placed \$1 million in the opaque box if they predicted that the player would One-Box, and put \$0 in the opaque box if they predicted that the player would Two-Box. There are two standard lines of argument about what the player should do. The first is a *causal dominance* argument which says that, because the player cannot cause money to be placed in the opaque box, they will always get at least as much money by taking both boxes than by taking one. The second is a *conditional expectation* argument which says that (because the predictor is highly reliable) One-Boxing provides strong evidence that there is \$1 million in the opaque box, and therefore the player should One-Box on the grounds that the conditional expected payoff given One-Boxing is higher than that of Two-Boxing. These are examples of *causal* and *evidential* decision-theoretic reasoning, respectively.

[28] Note that the little public discussion of FDT by academic philosophers has been largely critical (Schwarz, 2018; MacAskill, 2019).

and Soares (2017) refer to as *subjunctive* dependence. They explain this by stating that "When two physical systems are computing the same function, we will say that their behaviors "subjunctively depend" upon that function" (p. 6). Thus, in FDT, the expected utility given an action $a$ is computed by determining what the outcome of the decision problem would be if all relevant instances of the agent's decision-making algorithm output $a$.

In this section, we will assume an *a*causal stance on decision theory, that is, one other than CDT. There are several motivations for using a decision theory other than CDT:

- Intuitions about the appropriate decisions in thought experiments such as Newcomb's problem, as well as defenses of apparent failures of acausal decision theory in others (in particular, the "tickle defense" of evidential decision theory in the so-called smoking lesion case; see Ahmed (2014) for extensive discussion);
- Conceptual difficulties with causality (Schaffer, 2016);
- Demonstrations that agents using causal decision theory are exploitable in various ways (Kokotajlo, 2019b; Oesterheld and Conitzer, 2019);
- The *evidentialist wager* (MacAskill et al., 2019), which goes roughly as follows: In a large world (more below), we can have a far greater influence if we account for the acausal evidence our actions provide for the actions of others. So, under decision-theoretic uncertainty, we should wager in favor of decision theories which account for such acausal evidence.

We consider these sufficient motivation to study the implications of acausal decision theory for the reasoning of consequentialist agents. In particular, in this section we take up various possibilities for *acausal trade* between TAI systems. If we account for the evidence that one's choices provides for the choices that causally disconnected agents, this opens up both qualitatively new possibilities for interaction and quantitatively many more agents to interact with. Crucially, due to the potential scale of value that could be gained or lost via acausal interaction with vast numbers of distant agents, ensuring that TAI agents handle decision-theoretic problems correctly may be even more important than ensuring that they have the correct goals.

Agents using an acausal decision theory may coordinate in the absence of causal interaction. A concrete illustration is provided in Example 7.0.1, reproduced from Oesterheld (2017b)'s example, which is itself based on an example in Hofstadter (1983).

**Example 7.0.1** (Hofstadter's evidential cooperation game)**.** Hofstadter sends 20 participants the same letter, asking them to respond with a single letter 'C' (for cooperate) or 'D' (for defect) without communicating with each other. Hofstadter explains that by sending in 'C', a participant can increase everyone else's payoff by $\$2$. By sending in 'D', participants can increase their own payoff by $\$5$. The letter ends by informing the participants that they were all chosen for their high levels of rationality and correct decision making in weird scenarios like this. Note that every participant only cares about the balance of her own bank account and not about Hofstadter's or the other 19 participants'. Should you, as a participant, respond with 'C' or 'D'?

An acausal argument in favor of 'C' is: If I play 'C', this gives me evidence that the other participants also chose 'C'. Therefore, even though I cannot cause others to play 'C' — and therefore, on a CDT analysis — should play 'D' — the conditional expectation of my payoff given that I play 'C' is higher than my conditional expectation given that I play 'D'.

We will call this mode of coordination *evidential cooperation*.

For a satisfactory theory of evidential cooperation, we will need to make precise what it means for agents to be evidentially (but not causally) dependent. There are at least three possibilities.

1. Agents may tend to make the decisions on some reference class of decision problems. (That is, for some probability distribution on decision contexts $C$, $P(\text{Agent 1's decision in context } C = \text{Agent 2's decision in context } C)$ is high.)

2. An agent's taking action A in context C may provide evidence about the number of agents in the world who take actions like A in contexts like C.

3. If agents have similar source code, their decisions provide *logical* evidence for their counterpart's decision. (In turn, we would like a rigorous account of the notion of "source code similarity".)

It is plausible that we live in an infinite universe with infinitely many agents (Tegmark, 2003). In principle, evidential cooperation between agents in distant regions of the universe is possible; we may call this *evidential cooperation in large worlds (ECL)*.[29] If ECL were feasible then it is possible that it would allow agents to reap large amounts of value via acausal coordination. Treutlein (2019) develops a bargaining model of ECL and lists a number of open questions facing his formalism. Leskela (2019) addresses fundamental limitations on simulations as a tool for learning about distant agents, which may be required to gain from ECL and other forms of "acausal trade". Finally, Yudkowsky (n.d.) lists potential downsides to which agents may be exposed by reasoning about distant agents. The issues discussed by these authors, and perhaps many more, will need to be addressed in order to establish ECL and acausal trade as serious possibilities. Nevertheless, the stakes strike us as great enough to warrant further study.

[29] Oesterheld (2017b), who introduced the idea, calls this "multiverse-wide superrationality", following Hofstadter (1983)'s use of "superrational" to describe agents who coordinate acausally.

# 8 Acknowledgements

As noted in the document, several sections of this agenda were developed from writings by Lukas Gloor, Daniel Kokotajlo, Caspar Oesterheld, and Johannes Treutlein. Thank you very much to David Althaus, Tobias Baumann, Alexis Carlier, Alex Cloud, Max Daniel, Michael Dennis, Lukas Gloor, Adrian Hutter, Daniel Kokotajlo, János Kramár, David Krueger, Anni Leskelä, Matthijs Maas, Linh Chi Nguyen, Richard Ngo, Caspar Oesterheld, Mahendra Prasad, Rohin Shah, Carl Shulman, Stefan Torges, Johannes Treutlein, and Jonas Vollmer for comments on drafts of this document. Thank you also to the participants of the Center on Long-Term Risk research retreat and workshops, whose contributions also helped to shape this agenda.

# References

Arif Ahmed. *Evidence, decision and causality*. Cambridge University Press, 2014.

. AI Impacts. Likelihood of discontinuous progress around the development of agi. https://aiimpacts.org/likelihood-of-discontinuous-progress-around-the-development-of-agi/, 2018. Accessed: July 1 2019.

Riad Akrour, Marc Schoenauer, and Michele Sebag. Preference-based policy learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 12–27. Springer, 2011.

Steffen Andersen, Seda Ertaç, Uri Gneezy, Moshe Hoffman, and John A List. Stakes matter in ultimatum games. *American Economic Review*, 101(7):3427–39, 2011.

Giulia Andrighetto, Daniela Grieco, and Rosaria Conte. Fairness and compliance in the extortion game. 2015.

Scott E Atkinson, Todd Sandler, and John Tschirhart. Terrorism in a bargaining framework. *the Journal of Law and Economics*, 30(1):1–21, 1987.

Robert Axelrod. On six advances in cooperation theory. *Analyse & Kritik*, 22(1): 130–151, 2000.

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

Kyle Bagwell. Commitment and observability in games. *Games and Economic Behavior*, 8(2):271–280, 1995.

Tobias Baumann. Surrogate goals to deflect threats. http://s-risks.org/using-surrogate-goals-to-deflect-threats/, 2017. Accessed March 6, 2019.

Tobias Baumann. Challenges to implementing surrogate goals. http://s-risks.org/challenges-to-implementing-surrogate-goals/, 2018. Accessed March 6, 2019.

Tobias Baumann, Thore Graepel, and John Shawe-Taylor. Adaptive mechanism design: Learning to promote cooperation. *arXiv preprint arXiv:1806.04067*, 2018.

Ken Binmore, Ariel Rubinstein, and Asher Wolinsky. The nash bargaining solution in economic modelling. *The RAND Journal of Economics*, pages 176–188, 1986.

Iris Bohnet, Bruno S Frey, and Steffen Huck. More order with less law: On contract enforcement, trust, and crowding. *American Political Science Review*, 95(1):131–144, 2001.

Friedel Bolle, Yves Breitmoser, and Steffen Schlächter. Extortion in the laboratory. *Journal of Economic Behavior & Organization*, 78(3):207–218, 2011.

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

Nick Bostrom. Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pages 277–284, 2003.

Nick Bostrom. Superintelligence: paths, dangers, strategies. 2014.

Ronen I Brafman and Moshe Tennenholtz. Efficient learning equilibrium. In *Advances in Neural Information Processing Systems*, pages 1635–1642, 2003.

R. A. Briggs. Normative theories of rational choice: Expected utility. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.

Ernst Britting and Hartwig Spitzer. The open skies treaty. *Verification Yearbook*, pages 221–237, 2002.

Colin Camerer and Teck Hua Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999.

Colin F Camerer. *Behavioural game theory*. Springer, 2008.

Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

Christopher Cherniak. Computational complexity and the universal acceptance of logic. *The Journal of Philosophy*, 81(12):739–758, 1984.

Thomas J Christensen and Jack Snyder. Chain gangs and passed bucks: Predicting alliance patterns in multipolarity. *International organization*, 44(2):137–168, 1990.

Paul Christiano. Approval directed agents. https://ai-alignment.com/model-free-decisions-6e6609f5d99e, 2014. Accessed: March 15 2019.

Paul Christiano. Humans consulting hch. https://ai-alignment.com/humans-consulting-hch-f893f6051455, 2016a.

Paul Christiano. Prosaic ai alignment. https://ai-alignment.com/prosaic-ai-control-b959644d79c2, 2016b. Accessed: March 13 2019.

Paul Christiano. Clarifying "ai alignment". https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6, 2018a. Accessed: October 10 2019.

Paul Christiano. Preface to the sequence on iterated amplification. https://www.lesswrong.com/s/XshCxPjnBec52EcLB/p/HCv2uwgDGf5dyX5y6, 2018b. Accessed March 6, 2019.

Paul Christiano. Preface to the sequence on iterated amplification. https://www.lesswrong.com/posts/HCv2uwgDGf5dyX5y6/preface-to-the-sequence-on-iterated-amplification, 2018c. Accessed: October 10 2019.

Paul Christiano. Techniques for optimizing worst-case performance. https://ai-alignment.com/techniques-for-optimizing-worst-case-performance-39eafec74b99, 2018d. Accessed: June 24, 2019.

Paul Christiano. What failure looks like. https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like, 2019. Accessed: July 2 2019.

Paul Christiano and Robert Wiblin. Should we leave a helpful message for future civilizations, just in case humanity dies out? https://80000hours.org/podcast/episodes/paul-christiano-a-message-for-the-future/, 2019. Accessed: September 25, 2019.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

Mark Coeckelbergh. Can we trust robots? *Ethics and information technology*, 14(1): 53–60, 2012.

EA Concepts. Importance, tractability, neglectedness framework. https://concepts.effectivealtruism.org/concepts/importance-neglectedness-tractability/, n.d. Accessed: July 1 2019.

Ajeya Cotra. Iterated distillation and amplification. https://www.alignmentforum.org/posts/HqLxuZ4LhaFhmAHWk/iterated-distillation-and-amplification, 2018. Accessed: July 25 2019.

Jacob W Crandall, Mayada Oudah, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, Iyad Rahwan, et al. Cooperating with machines. *Nature communications*, 9(1):233, 2018.

Andrew Critch. A parametric, resource-bounded generalization of loeb's theorem, and a robust cooperation criterion for open-source game theory. *The Journal of Symbolic Logic*, pages 1–15, 2019.

Allan Dafoe. Ai governance: A research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 2018.

Wei Dai. Towards a new decision theory. https://www.lesswrong.com/posts/de3xjFaACCAk6imzv/towards-a-new-decision-theory, 2009. Accessed: March 5 2019.

Wei Dai. The main sources of ai risk. https://www.lesswrong.com/posts/WXvt8bxYnwBYpy9oT/the-main-sources-of-ai-risk, 2019. Accessed: July 2 2019.

Robyn M Dawes. Social dilemmas. *Annual review of psychology*, 31(1):169–193, 1980.

Karl W Deutsch and J David Singer. Multipolar power systems and international stability. *World Politics*, 16(3):390–406, 1964.

Daniel Dewey. My current thoughts on miri's "highly reliable agent design" work. https://forum.effectivealtruism.org/posts/SEL9PW8jozrvLnkb4/my-current-thoughts-on-miri-s-highly-reliable-agent-design, 2017. Accessed: October 6 2019.

Avinash Dixit. Trade expansion and contract enforcement. *Journal of Political Economy*, 111(6):1293–1317, 2003.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

K Eric Drexler. Reframing superintelligence: Comprehensive ai services as general intelligence, 2019.

Martin Dufwenberg and Uri Gneezy. Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, 30(2):163–182, 2000.

Daniel Ellsberg. The theory and practice of blackmail. Technical report, RAND CORP SANTA MONICA CA, 1968.

Johanna Etner, Meglena Jeleva, and Jean-Marc Tallon. Decision theory under ambiguity. *Journal of Economic Surveys*, 26(2):234–270, 2012.

Owain Evans, Andreas Stuhlmüller, Chris Cundy, Ryan Carey, Zachary Kenton, Thomas McGrath, and Andrew Schreiber. Predicting human deliberative judgments with machine learning. Technical report, Technical report, University of Oxford, 2018.

Tom Everitt, Jan Leike, and Marcus Hutter. Sequential extensions of causal and evidential decision theory. In *International Conference on Algorithmic DecisionTheory*, pages 205–221. Springer, 2015.

Tom Everitt, Daniel Filan, Mayank Daswani, and Marcus Hutter. Self-modification of policy and utility function in rational agents. In *International Conference on Artificial General Intelligence*, pages 1–11. Springer, 2016.

Tom Everitt, Pedro A Ortega, Elizabeth Barnes, and Shane Legg. Understanding agent incentives using causal influence diagrams, part i: single action settings. *arXiv preprint arXiv:1902.09980*, 2019.

James D Fearon. Rationalist explanations for war. *International organization*, 49(3): 379–414, 1995.

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868, 1999.

Ernst Fehr, Simon Gächter, and Georg Kirchsteiger. Reciprocity as a contract enforcement device: Experimental evidence. *ECONOMETRICA-EVANSTON ILL-*, 65:833–860, 1997.

Dan S Felsenthal and Abraham Diskin. The bargaining problem revisited: minimum utility point, restricted monotonicity axiom, and the mean as an estimate of expected utility. *Journal of Conflict Resolution*, 26(4):664–691, 1982.

Mark Fey and Kristopher W Ramsay. Mutual optimism and war. *American Journal of Political Science*, 51(4):738–754, 2007.

Mark Fey and Kristopher W Ramsay. Uncertainty and incentives in crisis bargaining: Game-free analysis of international conflict. *American Journal of Political Science*, 55(1):149–169, 2011.

Ben Fisch, Daniel Freund, and Moni Naor. Physical zero-knowledge proofs of physical properties. In *Annual Cryptology Conference*, pages 313–336. Springer, 2014.

Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

Lance Fortnow. Program equilibria and discounted computation time. In *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 128–133. ACM, 2009.

James W Friedman. A non-cooperative equilibrium for supergames. *The Review of Economic Studies*, 38(1):1–12, 1971.

Daniel Garber. Old evidence and logical omniscience in bayesian confirmation theory. 1983.

Ben Garfinkel. Revent developments in cryptography and possible long-run consequences. https://drive.google.com/file/d/0B0j9LKC65n09aDh4RmEzdlloT00/view, 2018. Accessed: November 11 2019.

Ben Garfinkel and Allan Dafoe. How does the offense-defense balance scale? *Journal of Strategic Studies*, 42(6):736–763, 2019.

Scott Garrabrant. Two major obstacles for logical inductor decision theory. https://agentfoundations.org/item?id=1399, 2017. Accessed: July 17 2019.

Scott Garrabrant and Abram Demski. Embedded agency. https://www.alignmentforum.org/posts/i3BTagvt3HbPMx6PN/embedded-agency-full-text-version, 2018. Accessed March 6, 2019.

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. Logical induction. *arXiv preprint arXiv:1609.03543*, 2016.

Alexandre Gazet. Comparative analysis of various ransomware virii. *Journal in computer virology*, 6(1):77–90, 2010.

Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

Allan Gibbard and William L Harper. Counterfactuals and two kinds of expected utility. In *Ifs*, pages 153–190. Springer, 1978.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18(2):141–153, 1989.

Alexander Glaser, Boaz Barak, and Robert J Goldston. A zero-knowledge protocol for nuclear warhead verification. *Nature*, 510(7506):497, 2014.

Charles L Glaser. The security dilemma revisited. *World politics*, 50(1):171–201, 1997.

Piotr J Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

Oded Goldreich and Yair Oren. Definitions and properties of zero-knowledge proof systems. *Journal of Cryptology*, 7(1):1–32, 1994.

Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM Journal on computing*, 18(1):186–208, 1989.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will ai exceed human performance? evidence from ai experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018.

Hilary Greaves, William MacAskill, Rossa O'Keeffe-O'Donovan, and Philip Trammell. Research agenda–web version a research agenda for the global priorities institute. 2019.

Avner Greif, Paul Milgrom, and Barry R Weingast. Coordination, commitment, and enforcement: The case of the merchant guild. *Journal of political economy*, 102(4): 745–776, 1994.

Frances S Grodzinsky, Keith W Miller, and Marty J Wolf. Developing artificial agents worthy of trust:"would you buy a used car from this artificial agent?". *Ethics and information technology*, 13(1):17–27, 2011.

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

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

Edward H Hagen and Peter Hammerstein. Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical population biology*, 69(3):339–348, 2006.

Joseph Y Halpern and Rafael Pass. Game theory with translucent players. *International Journal of Game Theory*, 47(3):949–976, 2018.

Lars Peter Hansen and Thomas J Sargent. *Robustness*. Princeton university press, 2008.

Lars Peter Hansen, Massimo Marinacci, et al. Ambiguity aversion and model misspecification: An economic perspective. *Statistical Science*, 31(4):511–515, 2016.

Garrett Hardin. The tragedy of the commons. *science*, 162(3859):1243–1248, 1968.

Paul Harrenstein, Felix Brandt, and Felix Fischer. Commitment and extortion. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 26. ACM, 2007.

John C Harsanyi and Reinhard Selten. A generalized nash solution for two-person bargaining games with incomplete information. *Management Science*, 18(5-part-2): 80–106, 1972.

Joseph Henrich, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, et al. Costly punishment across human societies. *Science*, 312(5781): 1767–1770, 2006.

Jack Hirshleifer. On the emotions as guarantors of threats and promises. *The Dark Side of the Force*, pages 198–219, 1987.

Douglas R Hofstadter. Dilemmas for superrational thinkers, leading up to a luring lottery. *Scientific American*, 6:267–275, 1983.

Terence Horgan. Counterfactuals and newcomb's problem. *The Journal of Philosophy*, 78(6):331–356, 1981.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*, pages 3326–3336, 2018.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Robert Jervis. Cooperation under the security dilemma. *World politics*, 30(2):167–214, 1978.

Robert Jervis. *Perception and Misperception in International Politics: New Edition*. Princeton University Press, 2017.

Daniel Kahneman, Ilana Ritov, David Schkade, Steven J Sherman, and Hal R Varian. Economic preferences or attitude expressions?: An analysis of dollar responses to public issues. In *Elicitation of preferences*, pages 203–242. Springer, 1999.

Ehud Kalai. Proportional solutions to bargaining situations: interpersonal utility comparisons. *Econometrica: Journal of the Econometric Society*, pages 1623–1630, 1977.

Ehud Kalai, Meir Smorodinsky, et al. Other solutions to nash's bargaining problem. *Econometrica*, 43(3):513–518, 1975.

Fred Kaplan. *The wizards of Armageddon*. Stanford University Press, 1991.

Holden Karnofsky. Some background on our views regarding advanced artificial intelligence. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence, 2016. Accessed: July 7 2019.

D Marc Kilgour and Frank C Zagare. Credibility, uncertainty, and deterrence. *American Journal of Political Science*, 35(2):305–334, 1991.

Stephen Knack and Philip Keefer. Institutions and economic performance: cross-country tests using alternative institutional measures. *Economics & Politics*, 7(3): 207–227, 1995.

Daniel Kokotajlo. The "commitment races" problem. https://www.lesswrong.com/posts/brXr7PJ2W4Na2EW2q/the-commitment-races-problem, 2019a. Accessed: September 11 2019.

Daniel Kokotajlo. Cdt agents are exploitable. *Unpublished working draft*, 2019b.

Peter Kollock. Social dilemmas: The anatomy of cooperation. *Annual review of sociology*, 24(1):183–214, 1998.

Kai A Konrad and Stergios Skaperdas. Credible threats in extortion. *Journal of Economic Behavior & Organization*, 33(1):23–39, 1997.

David M Kreps and Joel Sobel. Signalling. *Handbook of game theory with economic applications*, 2:849–867, 1994.

Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. Accountable algorithms. *U. Pa. L. Rev.*, 165:633, 2016.

David Krueger, Tegan Maharaj, Shane Legg, and Jan Leike. Misleading meta-objectives and hidden incentives for distributional shift. *Safe Machine Learning workshop at ICLR*, 2019.

Andrew Kydd. Which side are you on? bias, credibility, and mediation. *American Journal of Political Science*, 47(4):597–611, 2003.

Andrew H Kydd. Rationalist approaches to conflict prevention and resolution. *Annual Review of Political Science*, 13:101–121, 2010.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4190–4203, 2017.

Daryl Landau and Sy Landau. Confidence-building measures in mediation. *Mediation Quarterly*, 15(2):97–103, 1997.

Patrick LaVictoire, Benja Fallenstein, Eliezer Yudkowsky, Mihaly Barasz, Paul Christiano, and Marcello Herreshoff. Program equilibrium in the prisoner's dilemma via loeb's theorem. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*, 2019.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*, 2017.

Anni Leskela. Simulations as a tool for understanding other civilizations. Unpublished working draft, 2019.

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. *arXiv preprint arXiv:1811.08469*, 2018.

David Lewis. Prisoners' dilemma is a newcomb problem. *Philosophy & Public Affairs*, pages 235–240, 1979.

Xiaomin Lin, Stephen C Adams, and Peter A Beling. Multi-agent inverse reinforcement learning for certain general-sum stochastic games. *Journal of Artificial Intelligence Research*, 66:473–502, 2019.

Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

William MacAskill. A critique of functional decision theory. https://www.lesswrong.com/posts/ySLYSsNeFL5CoAQzN/a-critique-of-functional-decision-theory, 2019. Accessed: September 15 2019.

William MacAskill, Aron Vallinder, Caspar Oesterheld, Carl Shulman, and Johannes Treutlein. The evidentialist's wager. Manuscript, 2019.

Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6): 1447–1498, 2006.

Michael W Macy and Andreas Flache. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7229–7236, 2002.

Christopher JG Meacham. Binding and its consequences. *Philosophical studies*, 149 (1):49–71, 2010.

Kathleen L Mosier, Linda J Skitka, Susan Heers, and Mark Burdick. Automation bias: Decision making and performance in high-tech cockpits. *The International journal of aviation psychology*, 8(1):47–63, 1998.

Abhinay Muthoo. A bargaining model based on the commitment tactic. *Journal of Economic Theory*, 69:134–152, 1996.

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5):1313–1326, 1995.

John Nash. Two-person cooperative games. *Econometrica*, 21:128–140, 1953.

John F Nash. The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162, 1950.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

Douglass C North. Institutions. *Journal of economic perspectives*, 5(1):97–112, 1991.

Robert Nozick. Newcomb's problem and two principles of choice. In *Essays in honor of Carl G. Hempel*, pages 114–146. Springer, 1969.

Caspar Oesterheld. Deep reinforcement learning from human preferences. https://casparoesterheld.files.wordpress.com/2018/01/rldt.pdf, 2017a.

Caspar Oesterheld. Multiverse-wide cooperation via correlated decision making. 2017b.

Caspar Oesterheld. Robust program equilibrium. *Theory and Decision*, pages 1–17, 2019.

Caspar Oesterheld and Vincent Conitzer. Extracting money from causal decision theorists. 2019. Accessed: March 13 2019.

Stephen M Omohundro. The nature of self-improving artificial intelligence. *Singularity Summit*, 2008, 2007.

Stephen M Omohundro. The basic ai drives. In *AGI*, volume 171, pages 483–492, 2008.

OpenAI. Openai charter. https://openai.com/charter/, 2018. Accessed: July 7 2019.

Petro A Ortega and Vishal Maini. Building safe artificial intelligence: specification, robustness, and assurance. https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1, 2018. Accessed: July 7 2019.

Raja Parasuraman and Dietrich H Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410, 2010.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, pages 3643–3652, 2017.

Alexander Peysakhovich and Adam Lerer. Consequentialist conditional cooperation in social dilemmas with imperfect information. *arXiv preprint arXiv:1710.06975*, 2017.

Robert Powell. Bargaining theory and international conflict. *Annual Review of Political Science*, 5(1):1–30, 2002.

Robert Powell. War as a commitment problem. *International organization*, 60(1): 169–203, 2006.

Kai Quek. Rationalist experiments on war. *Political Science Research and Methods*, 5 (1):123–142, 2017.

Matthew Rabin. Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302, 1993.

Neil C Rabinowitz, Frank Perbet, H Francis Song, Chiyuan Zhang, SM Eslami, and Matthew Botvinick. Machine theory of mind. *arXiv preprint arXiv:1802.07740*, 2018.

Werner Raub. A general game-theoretic model of preference adaptations in problematic social situations. *Rationality and Society*, 2(1):67–93, 1990.

Robert W Rauchhaus. Asymmetric information, mediation, and conflict management. *World Politics*, 58(2):207–241, 2006.

Jonathan Renshon, Julia J Lee, and Dustin Tingley. Emotions and the micro-foundations of commitment problems. *International Organization*, 71(S1):S189–S218, 2017.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635, 2011.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109, 1982.

Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4):105–114, 2015.

Stuart J Russell and Devika Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 1994.

Santiago Sanchez-Pages. Bargaining and conflict with incomplete information. *The Oxford Handbook of the Economics of Peace and Conflict. Oxford University Press, New York*, 2012.

Wiliam Saunders. Hch is not just mechanical turk. https://www.alignmentforum.org/posts/4JuKoFguzuMrNn6Qr/hch-is-not-just-mechanical-turk?_ga=2.41060900.708557547.1562118039-599692079.1556077623, 2019. Accessed: July 2 2019.

Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242, 1999.

Jonathan Schaffer. The metaphysics of causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2016 edition, 2016.

James A Schellenberg. A comparative test of three models for solving "the bargaining problem". *Behavioral Science*, 33(2):81–96, 1988.

Thomas Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

David Schmidt, Robert Shupp, James Walker, TK Ahn, and Elinor Ostrom. Dilemma games: game parameters and matching protocols. *Journal of Economic Behavior & Organization*, 46(4):357–377, 2001.

Wolfgang Schwarz. On functional decision theory. umsu.de/wo/2018/688, 2018. Accessed: September 15 2019.

Anja Shortland and Russ Roberts. Shortland on kidnap. http://www.econtalk.org/anja-shortland-on-kidnap/, 2019. Accessed: July 13 2019.

Carl Shulman. Omohundro's "basic ai drives" and catastrophic risks. *Manuscript*, 2010.

Linda J Skitka, Kathleen L Mosier, and Mark Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999.

Alastair Smith and Allan C Stam. Bargaining and the nature of war. *Journal of Conflict Resolution*, 48(6):783–813, 2004.

Glenn H Snyder. " prisoner's dilemma" and" chicken" models in international politics. *International Studies Quarterly*, 15(1):66–103, 1971.

Nate Soares and Benja Fallenstein. Toward idealized decision theory. *arXiv preprint arXiv:1507.01986*, 2015.

Nate Soares and Benya Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, pages 103–125. Springer, 2017.

Joel Sobel. A theory of credibility. *The Review of Economic Studies*, 52(4):557–573, 1985.

Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.

Kaj Sotala. Disjunctive scenarios of catastrophic ai risk. In *Artificial Intelligence Safety and Security*, pages 315–337. Chapman and Hall/CRC, 2018.

Tom Florian Sterkenburg. The foudations of solomonoff prediction. Master's thesis, 2013.

Joerg Stoye. Statistical decisions under ambiguity. *Theory and decision*, 70(2):129–148, 2011.

Joseph Suarez, Yilun Du, Phillip Isola, and Igor Mordatch. Neural mmo: A massively multiagent game environment for training and evaluating intelligent agents. *arXiv preprint arXiv:1903.00784*, 2019.

Chiara Superti. Addiopizzo: Can a label defeat the mafia? *Journal of International Policy Solutions*, 11(4):3–11, 2009.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

William Talbott. Bayesian epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

Jessica Taylor. My current take on the paul-miri disagreement on alignability of messy ai. https://agentfoundations.org/item?id=1129, 2016. Accessed: October 6 2019.

Max Tegmark. Parallel universes. *Scientific American*, 288(5):40–51, 2003.

Moshe Tennenholtz. Program equilibrium. *Games and Economic Behavior*, 49(2): 363–373, 2004.

Johannes Treutlein. Modeling multiverse-wide superrationality. Unpublished working draft., 2019.

Jonathan Uesato, Ananya Kumar, Csaba Szepesvari, Tom Erez, Avraham Ruderman, Keith Anderson, Nicolas Heess, Pushmeet Kohli, et al. Rigorous agent evaluation: An adversarial approach to uncover catastrophic failures. *arXiv preprint arXiv:1812.01647*, 2018.

Eric Van Damme. The nash bargaining solution is optimal. *Journal of Economic Theory*, 38(1):78–100, 1986.

Hal R Varian. Computer mediated transactions. *American Economic Review*, 100(2): 1–10, 2010.

Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

Kenneth N Waltz. The stability of a bipolar world. *Daedalus*, pages 881–909, 1964.

Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach. *arXiv preprint arXiv:1803.00162*, 2018.

E Roy Weintraub. Game theory and cold war rationality: A review essay. *Journal of Economic Literature*, 55(1):148–61, 2017.

Sylvia Wenmackers and Jan-Willem Romeijn. New theory about old evidence. *Synthese*, 193(4):1225–1250, 2016.

Lantao Yu, Jiaming Song, and Stefano Ermon. Multi-agent adversarial inverse reinforcement learning. *arXiv preprint arXiv:1907.13220*, 2019.

Eliezer Yudkowsky. Ingredients of timeless decision theory. https://www.lesswrong.com/posts/szfxvS8nsxTgJLBHs/ingredients-of-timeless-decision-theory, 2009. Accessed: March 14 2019.

Eliezer Yudkowsky. Intelligence explosion microeconomics. *Machine Intelligence Research Institute, accessed online October*, 23:2015, 2013.

Eliezer Yudkowsky. Modeling distant superintelligences. https://arbital.com/p/distant_SIs/, n.d. Accessed: Feb. 6 2019.

Eliezer Yudkowsky and Nate Soares. Functional decision theory: A new theory of instrumental rationality. *arXiv preprint arXiv:1710.05060*, 2017.

Claire Zabel and Luke Muehlhauser. Information security careers for gcr reduction. https://forum.effectivealtruism.org/posts/ZJiCfwTy5dC4CoxqA/information-security-careers-for-gcr-reduction, 2019. Accessed: July 17 2019.

Chongjie Zhang and Victor Lesser. Multi-agent learning with policy prediction. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.