Servant of Many Masters: Shifting priorities in Pareto-optimal sequential decision-making

Andrew Critch, Stuart Russell
University of California, Berkeley
{critch, russell}@berkeley.edu

Abstract

It is often argued that an agent making decisions on behalf of two or more principals who have different utility functions should adopt a *Pareto-optimal* policy, i.e., a policy that cannot be improved upon for one agent without making sacrifices for another. A famous theorem of Harsanyi shows that, when the principals have a common prior on the outcome distributions of all policies, a Pareto-optimal policy for the agent is one that maximizes a fixed, weighted linear combination of the principals' utilities.

In this paper, we show that Harsanyi's theorem does not hold for principals with different priors, and derive a more precise generalization which does hold, which constitutes our main result. In this more general case, the relative weight given to each principal's utility should evolve over time according to how well the agent's observations conform with that principal's prior. The result has implications for the design of contracts, treaties, joint ventures, and robots.

1 Introduction

As AI systems take on an increasingly pivotal decision-making role in human society, an important question arises: *Whose values should a powerful decision-making machine be built to serve?* [Bostrom, 2014]

Consider, informally, a scenario wherein two or more principals—perhaps individuals, companies, or states—are considering cooperating to build or otherwise obtain an "agent" that will then interact with an environment on their behalf. The "agent" here could be anything that follows a policy, such as a robot, a corporation, or a web-based AI system. In such a scenario, the principals will be concerned with the question of "how much" the agent will prioritize each principal's interests, a question which this paper addresses quantitatively.

One might be tempted to model the agent as maximizing the expected value, given its observations, of some utility function U of the environment that equals a weighted sum

$$w^1 U^1 + w^2 U^2 \tag{1}$$

of the principals' individual utility functions U^1 and U^2, as Harsanyi's social aggregation theorem [Harsanyi, 1980] recommends. Then the question of prioritization could be reduced to that of choosing values for the weights w^i.

However, this turns out to be a suboptimal approach, from the perspective of the principals. As we shall see in Proposition 1, this solution form is not generally compatible with Pareto-optimality when agents have different beliefs. Harsanyi's setting does not account for agents having different priors, nor for decisions being made sequentially, after future observations.

In such a setting, we need a new form of solution, exhibited in this paper. The solution is presented along with a recursion (Theorem 3) that characterizes solutions by a process algebraically similar to, but meaningfully different from, Bayesian updating. The updating process resembles a kind of bet-settling between the principals, which allows them each to expect to benefit from the veracity of their own beliefs.

Qualitatively, this phenomenon can be seen in isolation whenever two people make a bet on a piece of decision-irrelevant trivia. If neither Alice nor Bob would base any important decision on whether Michael Jackson was born in 1958 or 1959, they might still make a bet for $100 on the answer. For a person chosen to arbitrate the bet (their "agent"), Michael Jackson's birth year now becomes a decision-relevant observation: it determines which of Alice and Bob gets the money!

Even in scenarios where differences in belief are not decision-irrelevant, once might expect some "degree" of bet-settling to arise from the disagreement. The main result of this paper (Theorem 3) is a precise formulation of exactly how and how much a Pareto-optimal agent will tend to prioritize each of its principals over time, as a result of differences in their implicit predictions about the agent's observations.

Related work

This paper may be viewed as extending or complimenting results in several areas:

Value alignment theory. The "single principal" value alignment problem—that of aligning the value function of an agent with the values of single human, or a team of humans in close agreement with one another—is already a very difficult one and should not be swept under the rug; approaches like inverse reinforcement learning (IRL) [Russell, 1998] [Ng *et al.*, 2000] [Abbeel and Ng, 2004] and cooperative inverse reinforcement learning (CIRL) [Hadfield-Menell *et al.*, 2016] have only begun to address it.

Social choice theory. The whole of social choice theory and voting theory may be viewed as an attempt to specify an agreeable formal policy to enact on behalf of a group. Harsanyi's utility aggregation theorem [Harsanyi, 1980] suggests one form of solution: maximizing a linear combination of group members' utility functions. The present work shows that this solution is inappropriate when principals have different beliefs, and Theorem 3 may be viewed as an extension of Harsanyi's form that accounts simultaneously for differing priors and the prospect of future observations. Indeed, Harsanyi's form follows as a direct corollary of Theorem 3 when principals do share the same beliefs (Corollary 4).

Bargaining theory. The formal theory of bargaining, as pioneered by [Nash, 1950] and carried on by [Myerson, 1979], [Myerson, 2013], and [Myerson and Satterthwaite, 1983], is also topical. Future investigation in this area might be aimed at generalizing their work to sequential decision-making settings, and this author recommends a focus on research specifically targeted at resolving conflicts.

Multi-agent systems. There is ample literature examining multi-agent systems using sequential decision-making models. Shoham and Leyton-Brown [2008] survey various models of multiplayer games using an MDP to model each agent's objectives. Chapter 9 of the same text surveys social choice theory, but does not account for sequential decision-making.

Zhang and Shah [2014] may be considered a sequential decision-making approach to social choice: they use MDPs to represent the decisions of players in a competitive game, and exhibit an algorithm for the players that, if followed, arrives at a Pareto-optimal Nash equilibrium satisfying a certain fairness criterion. Among the literature surveyed here, that paper is the closest to the present work in terms of its intended application: roughly speaking, achieving mutually desirable outcomes via sequential decision-making. However, that work is concerned with an ongoing interaction between the players, rather than selecting a policy for a single agent to follow as in this paper.

Multi-objective sequential decision-making. There is also a good deal of work on Multi-Objective Optimization (MOO) [Tzeng and Huang, 2011], including for sequential decision-making, where solution methods have been called Multi-Objective Reinforcement Learning (MORL). For instance, Gábor *et al.* [1998] introduce a MORL method called Pareto Q-learning for learning a set of a Pareto-optimal polices for a Multi-Objective MDP (MOMDP). Soh and Demiris [2011] define Multi-Reward Partially Observable Markov Decision Processes (MR-POMDPs), and use use genetic algorithms to produce non-dominated sets of policies for them. Roijers *et al.* [2015] refer to the same problems as Multi-objective POMDPS (MOPOMDPs), and provide a bounded approximation method for the optimal solution set for all possible weightings of the objectives. Wang [2014] surveys MORL methods, and contributes Multi-Objective Monte-Carlo Tree Search (MOMCTS) for discovering multiple Pareto-optimal solutions to a multi-objective optimization problem. Wray and Zilberstein [2015] introduce Lexicographic Partially Observable Markov Decision Process (LPOMDPs), along with two accompanying solution methods.

However, none of these or related works addresses scenarios where the objectives are derived from principals with differing beliefs, from which the priority-shifting phenomenon of Theorem 3 arises. Differing beliefs are likely to play a key role in negotiations, so for that purpose, the formulation of multi-objective decision-making adopted here is preferable.

2 Notation

Random variables are denoted by uppercase letters, e.g., S_1, and lowercase letters, e.g., s_1, are used as indices

ranging over the values of a variable, as in the equation

$$\mathbb{E}[S_1] = \sum_{s_1} \mathbb{P}(s_1) \cdot s_1.$$

Given a set A, the set of probability distributions on A is denoted ΔA.

Sequences are denoted by overbars, e.g., given a sequence $(s_1, \ldots, s_n)$, $\bar{s}$ stands for the whole sequence. Subsequences are denoted by subscripted inequalities, so e.g., $s_{\leq 4}$ stands for (s_1, s_2, s_3), and $s_{>4}$ stands for $(s_5, \ldots, s_n)$.

3 Formalism

N.B.: All results in this paper generalize directly from agents with two principals to agents with several, but for clarity of exposition, the case of two principals will be prioritized.

Consider a scenario wherein Alice and Bob will share some cake, and have different predictions of the cake's color. Even if the color would be decision-irrelevant for either Alice or Bob on their own (they don't care what color the cake is), we will show that the difference between their predictions will tend to make the cake color a decision-relevant observation for a Pareto-optimal cake-splitting policy that is adopted before they see the cake. Specifically, we will show that Pareto-optimal policies tend to incorporate some degree of bet-settling between Alice and Bob, where the person who was more right about the color of the cake will end up getting more of it.

Serving multiple principals as a single POMDP

To formalize such scenarios, where a single agent acts on behalf of multiple principals, we need some definitions.

We encode each principal j's view of the agent's decision problem as a finite horizon POMDP, $D^j = (\mathcal{S}^j, \mathcal{A}, T^j, U^j, \mathcal{O}, \Omega^j, n)$, which simultaneously represents that principal's beliefs about the environment, and the principal's utility function (see Russell *et al.* [2003] for an introduction to POMDPs). These symbols take on their usual meaning:

- $\mathcal{S}^j$ represents a set of possible states s of the environment,
- $\mathcal{A}$ represents the set of possible actions a available to the agent,
- T^j represents the conditional probabilities principal j believes will govern the environment state transitions, i.e., $\mathbb{P}^j(s_{i+1} \mid s_i a_i)$,

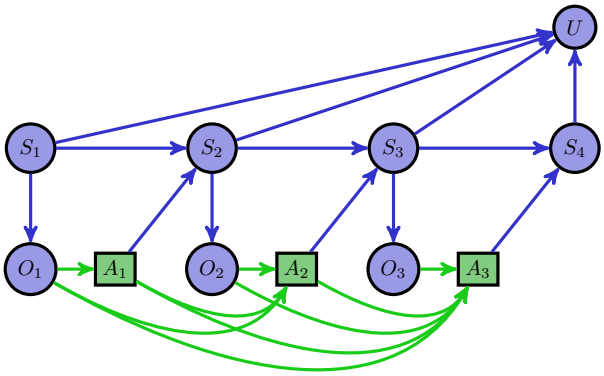

Figure 1: A POMDP with horizon $n = 3$ (in blue), being solved by a full-memory policy (in green).

- U^j represents principal j's utility function from sequences of environmental states $(s_1, \ldots, s_n)$ to $\mathbb{R}$; for the sake of generality, U^j is *not assumed* to be additive over time, as reward functions often are,
- $\mathcal{O}$ represents the set of possible observations o of the agent,
- Ω^j represents the conditional probabilities principal j believes will govern the agent's observations, i.e., $\mathbb{P}^j(o_i \mid s_i)$, and
- n is the horizon (number of time steps)

This POMDP structure is depicted by the Bayesian network in Figure 1. (See Darwiche [2009] for an intro to Bayesian networks.) At each point in time i, the agent has a time-specific policy π_i, which receives the agent's *history*,

$$h_i := (o_{\leq i}, a_{<i}),$$

and returns a distribution $\pi_i(- \mid h_i)$ on actions a_i, which will then be used to generate an action a_i with probability $\pi(a_i \mid h_i)$. Thus, principal j's subjective probability of an outcome $(\bar{s}, \bar{o}, \bar{a})$ is given by a probability distribution $\mathbb{P}^j$ that takes π as a parameter:

$$\mathbb{P}^j(\bar{s}, \bar{o}, \bar{a}; \pi) := \mathbb{P}^j(s_1) \cdot \prod_{i=1}^{n} \mathbb{P}^j(o_i \mid s_i) \, \pi(a_i \mid h_i) \, \mathbb{P}^j(s_{i+1} \mid s_i, a_i) \quad (2)$$

Full-memory assumption. *Every policy π in this paper will be assumed to employ a "full memory"*, so it decomposes into a sequence of policies π_i for each time step. In Figure 1, the part of the Bayes net governed by the full-memory policy is highlighted in green.

Common knowledge assumptions. It is assumed that the principals will have common knowledge of the (full-memory) policy $\pi = (\pi_1, \ldots, \pi_n)$ they select for the agent to implement, but that the principals may have different beliefs about how the environment works, and of course different utility functions. It is also assumed that the principals have common knowledge of one another's current beliefs at the time of the agent's creation, which we refer to as their their priors.

This last assumption is critical. During the agent's creation, one should expect each principal's beliefs to have updated somewhat in response to disagreements from the other. Assuming common knowledge of their priors means assuming the principals to have reached an equilibrium where, each knowing what the other believes, they do not wish to further update their own beliefs.[1]

Pareto-optimal policies

A policy will be considered Pareto-optimal relative to a set of POMDPs it could be deployed to solve.

Definition 1 (Compatible POMDPs)**.** *We say that two POMDPs, D^1 and D^2, are* compatible *if any policy for one may be viewed as a policy for the other, i.e., they have the same set of actions $\mathcal{A}$ and observations $\mathcal{O}$, and the same number of time steps n.*

In this context, where a single policy π may be evaluated relative to more than one POMDP, we use superscripts to represent which POMDP is governing the probabilities and expectations, e.g.,

$$\mathbb{E}^j[U^j; \pi] := \sum_{\bar{s} \in (\mathcal{S}^j)^n} \mathbb{P}^j(\bar{s}; \pi) U^j(\bar{s})$$

represents the expectation in D^j of the utility function U^j, assuming policy π is followed.

Definition 2 (Pareto-optimal policies)**.** *A policy π is* Pareto-optimal *for a set of compatible POMDPs $(D^1, \ldots, D^k)$ if for any other policy π' and any $j \in \{1, \ldots, k\}$*

$$\mathbb{E}^j[U^j; \pi'] > \mathbb{E}^j[U^j; \pi] \Rightarrow (\exists \ell) \left(\mathbb{E}^\ell[U^\ell; \pi'] < \mathbb{E}^\ell[U^\ell; \pi] \right),$$

It is assumed that, before the agent's creation, the principals will be seeking a Pareto-optimal (full-memory) policy for the agent to follow, relative to the POMDPs D^j describing each principal's view of the agent's task.

[1] It is enough to assume the principals have reached a "persistent disagreement" that cannot be mediated by the agent in some way. Future work should design solutions for facilitating the process of attaining common knowledge, or to obviate the need to assume it.

Example: cake betting

A quantitative model of a cake betting scenario is laid out in Table 1, and described as follows.

Alice (Principal 1) and Bob (Principal 2) are about to be presented with a cake which they can choose to split in half to share, or give entirely to one of them. They have (built or purchased) a robot that will make the cake-splitting decision on their behalf. Alice's utility function returns 0 if she gets no cake, 20 if she gets half a cake, or 30 if she gets a whole cake. Bob's utility function values Bob getting cake in the same way.

However, Alice and Bob have different beliefs about the color of the cake. Alice is 90% sure that the cake is red ($S_1 = O_1 =$ "red"), versus 10% sure it will be green ($S_1 = O_1 =$ "green"), whereas Bob's probabilities are reversed.

Upon seeing the cake, the robot must decide to either give Alice the entire cake ($A_1 = S_2 =$ (all, none)), split the cake half-and-half ($A_1 = S_2 =$ (half, half)), or give Bob the entire cake ($A_1 = S_2 =$ (none, all)). Moreover, Alice and Bob have common knowledge of all these facts.

Now, consider the following Pareto-optimal full-memory policy that favors Alice (Principal 1) when O_1 is red, and Bob (Principal 2) when O_1 is green:

$$\hat{\pi}(- \mid \text{red}) = 100\%(\text{all, none})$$
$$\hat{\pi}(- \mid \text{green}) = 100\%(\text{none, all})$$

This policy can be viewed intuitively as a bet between Alice and Bob about the value of O_1, and is highly appealing to both principals:

$$\mathbb{E}^1[U^1; \hat{\pi}] = 90\%(30) + 10\%(0) = 27$$
$$\mathbb{E}^2[U^2; \hat{\pi}] = 10\%(0) + 90\%(30) = 27$$

In particular, $\hat{\pi}$ is more appealing to both Alice and Bob than an agreement to deterministically split the cake (half, half), which would yield them each an expected utility of 20. However,

Proposition 1. *The Pareto-optimal strategy $\hat{\pi}$ above cannot be implemented by any agent that naïvely maximizes a fixed-over-time linear combination of the conditionally expected utilities of the two principals. That is, it cannot be implemented by any policy π satisfying*

$$\pi(- \mid o_1) \in \operatorname*{argmax}_{\alpha \in \Delta A} \Big(r \cdot \mathbb{E}^1[U^1 \mid o_1; a_1 \sim \alpha] + (1 - r) \cdot \mathbb{E}^2[U^2 \mid o_1; a_1 \sim \alpha] \Big) \quad (3)$$

for some fixed $r \in [0, 1]$. Moreover, every such policy π is strictly worse than $\hat{\pi}$ in expectation to one of the principals.

$S_1 = O_1$	$\mathbb{P}^1(O_1)$	$\mathbb{P}^2(O_1)$	$A_1 = S_2$	U^1	U^2
red cake	90%	10%	(all, none)	30	0
			(half, half)	20	20
			(none, all)	0	30
green cake	10%	90%	(all, none)	30	0
			(half, half)	20	20
			(none, all)	0	30

Table 1: An example scenario wherein a Pareto-optimal full-memory policy undergoes priority shifting (who gets the cake), based on features that are decision-irrelevant for each principal (cake color).

Proof. See appendix. □

This proposition is relatively unsurprising when one considers the full-memory policy $\hat{\pi}$ intuitively as a bet-settling mechanism, because the nature of betting is to favor different preferences based on future observations. However, to be sure of this impossibility claim, one must rule out the possibility that the $\hat{\pi}$ could be implemented by having the agent choose which element of the argmax in Equation 3 to use based on whether the cake appears red or green. (See appendix.)

Characterizing Pareto-optimality geometrically

With the definitions above, we can characterize a Pareto-optimality as a geometric condition.

Policy mixing assumption. Given policies $\pi^1, \ldots, \pi^R$ and a distribution $\alpha = (\alpha^1, \ldots, \alpha^R) \in \Delta\{1, \ldots, R\}$, we assume that the agent may construct a new policy by choosing at time 0 between the π^r with probability α^r, and then executing the chosen policy for the rest of time. We write this policy as $\pi = \sum_r \alpha^r \pi^r$, whence we derive:

$$\mathbb{E}^j\left[U^j; \sum_r \alpha^r \pi^r\right] = \sum_r \alpha^r \mathbb{E}^j[U^j; \pi^r]. \qquad (4)$$

Lemma 1 (Polytope Lemma)**.** *A full-memory policy π is Pareto-optimal to principals 1 and 2 if and only if there exist weights $w^1, w^2 \geq 0$ with $w^1 + w^2 = 1$ such that*

$$\pi \in \operatorname*{argmax}_{\pi^* \in \Pi} \left(w^1 \mathbb{E}^1[U^1; \pi^*] + w^2 \mathbb{E}^2[U^2; \pi^*]\right) \qquad (5)$$

Proof. The mixing assumption gives the set of policies Π the structure of a convex space that the maps $\mathbb{E}^j[U^j; -]$ respect by Equation 4. This ensures that the image of the map $f : \Pi \to \mathbb{R}^2$ given by

$$f(\pi) := \left(\mathbb{E}^1[U^1; \pi],\ \mathbb{E}^2[U^2; \pi]\right)$$

is a closed, convex polytope. As such, a point (x, y) lies on the Pareto boundary of $\operatorname{image}(f)$ if and only if there exist nonnegative weights (w^1, w^2), not both zero, such that

$$(x, y) \in \operatorname*{argmax}_{(x^*, y^*) \in \operatorname{image}(f)} \left(w^1 x^* + w^2 y^*\right)$$

After normalizing $w^1 + w^2$ to equal 1, this implies the result. □

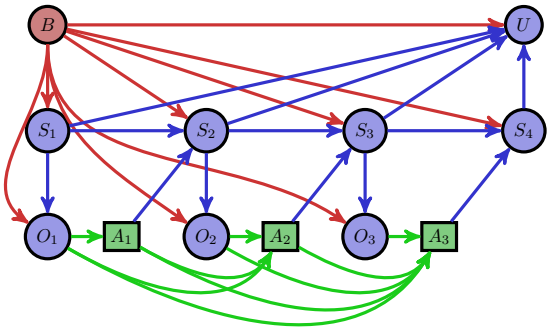

Figure 2: A POMDP (mixture) with horizon $n = 3$ initialized by a Boolean B, being solved by a full-memory policy (green)

Characterizing Pareto-optimality probabilistically

To help us apply the Polytope Lemma, we will adopt an interpretation wherein the weights w^i are subjective probabilities for the agent, as follows.

For any $w \in \Delta\{1, 2\}$, we define a new POMDP, D, that works by flipping a (w^1, w^2)-weighted coin, and then running D^1 or D^2 thereafter, according to the coin flip. We denote this by

$$D = w^1 D^1 + w^2 D^2,$$

and call D a *POMDP mixture*. A formal definition of D is given in the appendix. It can be depicted by a Bayes net by adding an additional environmental node for B in the diagram of D^1 and D^2 (see Figure 2).

Given any full-memory policy π, the expected payoff of

π in $w^1 D^1 + w^2 D^2$ is exactly

$$\begin{aligned} &\mathbb{P}(B=1)\cdot\mathbb{E}[U \mid B=1;\pi] \\ &+\mathbb{P}(B=2)\cdot\mathbb{E}[U \mid B=2;\pi] \\ &= w^1\mathbb{E}^2[U^1;\pi] + w^2\mathbb{E}^2[U^2;\pi] \end{aligned}$$

Therefore, using the above definitions, Lemma 1 may be restated in the following equivalent form:

Lemma 2 (Mixture Lemma)**.** *Given a pair (D^1, D^2) of compatible POMDPs, a full-memory policy π is Pareto-optimal for that pair if and only if there exists $w \in \Delta\{1,2\}$ such that π is an optimal full-memory policy for the single POMDP given by $w^1 D^1 + w^2 D^2$.*

Expressed in the form of Equation 5, it might not be clear how a Pareto-optimal full-memory policy makes use of its observations over time, aside from storing them in memory. For example, is there any sense in which the agent carries "beliefs" about the environment that it "updates" at each time step? Lemma 2 allows us to reduce some such questions about Pareto-optimal policies to questions about single POMDPs.

If π is an optimal full-memory policy for a single POMDP, the optimality of each action distribution $\pi_i(- \mid h_i)$ can be characterized without reference to the previous policy components $(\pi_1, \ldots, \pi_{i-1})$, nor to $\pi_i(- \mid h_i')$ for any alternate history h_i'. This can be expressed using Pearl's "Do()" notation [Pearl, 2009]:

Definition 3 ("do" notation)**.** *The probability of $\bar{o}$* causally conditioned on *$\bar{a}$ is defined as*

$$\begin{aligned} &\mathbb{P}^j(\bar{o} \mid \mathrm{Do}(\bar{a})) \\ &:= \sum_{\bar{s}\in(\mathcal{S}^j)^n} \mathbb{P}^j(s_1)\cdot\prod_{i=1}^{n}\mathbb{P}^j(o_i \mid s_i)\,\mathbb{P}^j(s_{i+1} \mid s_i a_i) \end{aligned}$$

Definition 4 (Expected utility abbreviation)**.** *For brevity, given any POMDP D and policy π, we write*

$$E_\pi^D(\alpha; h_i) := \mathbb{E}[U \mid h_i;\ a_n \sim \alpha;\ \pi_{>i}].$$

i.e., the total expected utility in D that would result from replacing $\pi_i(- \mid h_i)$ by α. This quantity does not depend on $\pi_{\leq i}$.

Proposition 2 (Classical separability)**.** *If D is a POMDP described by conditional probabilities $\mathbb{P}(- \mid -)$ and utility function U (as in Equation 2), then a full-memory policy π is optimal for D if and only if for each time step i and each observation/action history h_i, the action distribution $\pi_i(- \mid h_i)$ satisfies the following backward recursion:*

$$\pi_i(- \mid h_i) \in \operatorname*{argmax}_{\alpha\in\Delta A}\left(\mathbb{P}\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right)\cdot E_\pi^D(\alpha; h_i)\right)$$

This characterization of $\pi_i(-|h_i)$ does not refer to $\pi_1, \ldots, \pi_{i-1}$, nor to $\pi_i(h_i')$ for any alternate history h_i'.

Proof. This is just Bellman's Principle of Optimality. See [Bellman, 1957], Chap. III. 3. □

N.B.: Unlike Bellman's "backup" equation, the above proposition requires no assumption whatsoever on the form of the utility function. Note also that when the probability term $\mathbb{P}(o_{\leq i} \mid \mathrm{Do}(a_{<i}))$ is non-zero, it may be removed from the argmax *without changing the theorem statement. But when the term is zero, its presence is essential, and implies that $\pi_i(- \mid h_i)$ can be anything.*

It turns out that Pareto-optimality can be characterized in a similar way by backward recursion from the final time step. The resulting recursion reveals a pattern in how the weights on the principals' conditionally expected utilities must change over time, which is the main result of this paper:

Theorem 3 (Pareto-optimal control theorem)**.** *Given a pair (D^1, D^2) of compatible POMDPs with horizon n, a full-memory policy π is Pareto-optimal if and only if its components π_i for $i \leq n$ satisfy the following backward recursion for some weights $w \in \Delta\{1,2\}$:*

$$\begin{aligned} \pi^i(- \mid h_i) \in \operatorname*{argmax}_{\alpha\in\Delta A}\Bigg(&\\ w^1\mathbb{P}^1\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right)\cdot E_\pi^{D^1}(\alpha; h_i) &\\ + w^2\mathbb{P}^2\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right)\cdot E_\pi^{D^2}(\alpha; h_i)\Bigg) & \end{aligned}$$

In words, to achieve Pareto-optimality, the agent must

1. *use each principal's own world-model D^j when estimating the degree $E_\pi^{D^j}(\alpha; h_i)$ to which a decision α favors that principal's utility function, and*
2. *shift the relative priority of each principal's expected utility in the agent's maximizationtarget over time, by a factor proportional to how well that principal's prior predicts the agent's observations, $\mathbb{P}^i\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right)$.*

N.B.: The analogous result for more than two POMDPs holds as well, with essentially the same proof.

Proof of Theorem 3. By Lemma 2, the Pareto-optimality of π for (D^1, D^2) is equivalent to its classical optimality for $D = w^1 D^1 + w^2 D^2$ for some (w^1, w^2). Writing $\mathbb{P}$ for probabilities in D, Proposition 2 says this is equivalent

to $\alpha = \pi^i(- \mid h_i)$ maximizing the following expression $F(\alpha)$ for each i:

$$F(\alpha) = \mathbb{P}\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right) \cdot E_\pi^D(\alpha; h_i). \qquad (6)$$

The expectation factor on the right equals

$$\begin{aligned} E_\pi^D(\alpha; h_i) = \mathbb{P}\left(B = 1 \mid o_{\leq i}, \mathrm{Do}(a_{<i})\right) \cdot E_\pi^{D^1}(\alpha; h_i) \\ + \mathbb{P}\left(B = 2 \mid o_{\leq i}, \mathrm{Do}(a_{<i})\right) \cdot E_\pi^{D^2}(\alpha; h_i). \end{aligned}$$

Multiplying by

$$\begin{aligned} \mathbb{P}\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right) = w^1\mathbb{P}^1\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right) \\ + w^2\mathbb{P}^2\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right) \end{aligned}$$

and applying Bayes' rule yields that

$$\begin{aligned} F(\alpha) = w^1\mathbb{P}^1\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right) E_\pi^{D^1}(\alpha; h_i) \\ + w^2\mathbb{P}^2\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right) E_\pi^{D^2}(\alpha; h_i), \end{aligned}$$

hence the result. □

To see the necessity of the $\mathbb{P}^j$ terms that shift the expectation weights in Theorem 3 over time, recall from Proposition 1 that, without these, some Pareto-optimal policies cannot be implemented. These $\mathbb{P}^j$ terms are responsible for the "bet-settling" phenomena discussed in the introduction.

However, when the principals have the same beliefs, they aways assign the same probability to the agent's observations, so the weights on their respective valuations do not change over time. Hence, as a special instance, we derive:

Corollary 4 (Harsanyi's utility aggregation formula)**.** *Suppose that principals 1 and 2 share the same beliefs about the environment, i.e., the pair (D^1, D^2) of compatible POMDPs agree on all parameters except the principals' utility functions $U^1 \neq U^2$. Then a full-memory policy π is Pareto-optimal if and only if there exists $w \in \Delta\{1, 2\}$ such that for $i \leq n$, π_i satisfies*

$$\begin{aligned} \pi^i(- \mid h_i) \in \operatorname*{argmax}_{\alpha \in \Delta A} \big(\\ \mathbb{E}[w^1U^1 + w^2U^2] \mid h_i;\ a_i \sim \alpha;\ \pi_{>i}] \Big) \end{aligned}$$

where $\mathbb{E} = \mathbb{E}^1 = \mathbb{E}^2$ denotes the shared expectations of both principals.

Proof. Setting $\mathbb{E} = \mathbb{E}^1 = \mathbb{E}^2$ in Theorem 3, factoring out the common coefficient $\mathbb{P}^1\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right) = \mathbb{P}^2\left(o_{\leq i} \mid \mathrm{Do}(a_{<i})\right)$, and applying linearity of expectation yields the result. □

4 Conclusion

Theorem 3 exhibits a novel form for the objective of a sequential decision-making policy that is Pareto-optimal according to principals with differing beliefs.

This form represents two departures from naïve utility aggregation: to achieve Pareto-optimality for principals with differing beliefs, an agent must (1) use each principal's own beliefs (updated on the agent's observations) when evaluating how well an action will serve that principal's utility function, and (2) shift the relative priority it assigns to each principal's expected utilities over time, by a factor proportional to how well that principal's prior predicts the agent's observations.

Implications for contract design

Theorem 3 has implications for modeling and structuring the process of contract design. If a contract is being created between principals with different beliefs, then to the extent that the principals will target Pareto-optimality among them as an objective, there will be a tendency for the contract to end up implicitly settling bets between the principals. Perhaps making the bet-settling nature of Pareto-optimal contract design more explicit might help to design contracts that are more attractive to both principals, along the lines illustrated by Proposition 1. This could potentially lead to more successful negotiations, provided the principals remained willing to uphold the contract after its implicit bets have been settled.

Implications for shareable AI systems

Proposition 1 shows how the Pareto-optimal form of Theorem 3 is more attractive—from the perspective of the principals—than policies that do not account for differences in their beliefs. The relative attractiveness of shared ownership versus individual ownership of AI systems may be essential to the technological adoption of shared systems. Consider the following product substitutions that might be enabled by the development of shareable machine learning systems:

- Office assistant software jointly controlled by a team, as an improvement over personal assistant software for each member of the team.
- A team of domestic robots controlled by a family, as an improvement over individual robots each controlled by a separate family member.
- A web-based security system shared by several interested companies or nations, as an improvement over individual security systems deployed by each group.

It may represent a significant technical challenge for any of these substitutions to become viable. However, machine learning systems that are able to approximate Pareto-optimality as an objective are more likely to be sufficiently appealing to motivate the switch from individual control to sharing.

Implications for bargaining versus racing

Consider two nations—allies or adversaries—who must decide whether to cooperate in the deployment of a very powerful and autonomous AI system.

If the nations cannot reach agreement as to what policy a jointly owned AI system should follow, joint ownership may be less attractive than building separate AI systems, one for each party. This could lead to an arms race between nations competing under time pressure to develop ever more powerful militarized AI systems. Under such race conditions, everyone loses, as each nation is afforded less time to ensure the safety and value alignment of its own system.

The first author's primary motivation for this paper is to initiate a research program with the mission of averting such scenarios. Beginning work today on AI architectures that are more amenable to joint ownership could help lead to futures wherein powerful entities are more likely to share and less likely to compete for the ownership of such systems.

Future work

Insofar as Theorem 3 is not particularly mathematically sophisticated—it employs only basic facts about convexity and linear algebra—this suggests there may be more low-hanging fruit to be found in the domain of "machine implementable social choice theory". Future work should address methods for helping the principals to share information—perhaps in exchange for adjustments to the weights in Theorem 3—to reach either a state of agreement or a persistent disagreement that allows the theorem to be applied. More ambitiously, bargaining models that account for a degree of transparency between the principals should be employed, as individual humans and institutions have some capacity for detecting one another's intentions.

As well, scenarios where the principals continue to exhibit some active control over the system after its creation should be modeled in detail. In real life, principals usually continue to exist in their agents' environments, and accounting for this will be a separate technical challenge.

As a final motivating remark, consider that social choice theory and bargaining theory were both pioneered during the Cold War, when it was particularly compelling to understand the potential for cooperation between human institutions that might behave competitively. In the coming decades, machine intelligence will likely bring many new challenges for cooperation, as well as new means to cooperate, and new reasons to do so. As such, new technical aspects of social choice and bargaining will likely continue to emerge.

5 Appendix

Here we make available the technical details for defining POMDP mixtures, and proving that certain Pareto-optimal expectations cannot be obtained without priority-shifting.

Definition 5 (POMDP mixtures)**.** *Suppose that D^1 and D^2 are compatible POMDPs, with parameters $D^j = (\mathcal{S}^j, \mathcal{A}, T^j, U^j, \mathcal{O}, \Omega^j, n)$. Define a new POMDP compatible with both, denoted $D = w^1 D^1 + w^2 D^2$, with parameters $D^j = (\mathcal{S}, \mathcal{A}, T, U, \mathcal{O}, \Omega, n)$, as follows:*

- $\mathcal{S} := \{(j, s) \mid j \in \{1, 2\}, s \in \mathcal{S}^j\}$,
- *Environmental transition probabilities T given by*

$$\mathbb{P}\left((j, s_1)\right) := w^j \cdot \mathbb{P}^j(s_1)$$

for any initial state $s_1 \in \mathcal{S}^j$, and thereafter,

$$\mathbb{P}\left((j', s_{i+1}) \mid (j, s_i), a_i\right) := \begin{cases} \mathbb{P}^j\left(s_{i+1} \mid s_i a_i\right) & \text{if } j' = j \\ 0 & \text{if } j' \neq j \end{cases}$$

Hence, the value of j will be constant over time, so a full history for the environment may be represented by a pair

$$(j, \bar{s}) \in \{1\} \times (\mathcal{S}^1)^n \cup \{2\} \times (\mathcal{S}^2)^n.$$

Let B denote the boolean random variable that equals whichever constant value of j obtains, so then

$$\mathbb{P}(B = j) = w^j$$

- *The utility function U is given by*

$$U(j, \bar{s}) := U^j(\bar{s})$$

- *The observation probabilities Ω are given by*

$$\mathbb{P}\left(o_i \mid (j, s_i)\right) := \mathbb{P}(B = j) \cdot \mathbb{P}^j(o_i \mid s_i)$$

In particular, the agent does not observe directly whether $j = 1$ or $j = 2$.

Proof of Proposition 1. Suppose π is any policy satisfying Equation 3 for some fixed r, and consider the following cases for r:

1. If $r < 1/3$, then π must satisfy

   $$\pi(- \mid o_1) = 100\%\text{(none, all)}.$$

   Here, $\mathbb{E}^1[U^1; \pi] = 0 < 27$, so π is strictly worse than $\hat{\pi}$ in expectation to Alice.

2. If $r = 1/3$, then π must satisfy

   $$\pi(- \mid o_1) = q(o_1)\text{(none, all)}+(1-q(o_1))\text{(half, half)}$$

   for some $q(o_1) \in [0, 1]$ depending on o_1. Here, $\mathbb{E}^1[U^1; \pi] \leq 20 < 27$ (with equality when $q(\text{red}) = q(\text{green}) = 1$), so π is strictly worse than $\hat{\pi}$ in expectation to Alice.

3. If $1/3 < r < 2/3$, then π must satisfy

   $$\pi(- \mid o_1) = 100\%\text{(half, half)}$$

   Here, $\mathbb{E}^1[U^1; \pi] = \mathbb{E}^2[U^2; \pi] = 20 < 27$, so π is strictly worse than $\hat{\pi}$ in expectation to both Alice and Bob.

The remaining cases, $r = 2/3$ and $r > 2/3$, are symmetric to the first two, with Bob in place of Alice and (none, all) in place of (all, none).

Hence, no fixed linear combination of the principals' utility functions can be maximized to simultaneously achieve an expected utility of 27 for both players. □

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ., 1957.

Nick Bostrom. *Superintelligence: Paths, dangers, strategies*. OUP Oxford, 2014.

Adnan Darwiche. *Modeling and reasoning with Bayesian networks (Chapter 4)*. Cambridge University Press, 2009.

Zoltán Gábor, Zsolt Kalmár, and Csaba Szepesvári. Multi-criteria reinforcement learning. In *ICML*, volume 98, pages 197–205, 1998.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning, 2016.

John C Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. In *Essays on Ethics, Social Behavior, and Scientific Explanation*, pages 6–23. Springer, 1980.

Roger B Myerson and Mark A Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of economic theory*, 29(2):265–281, 1983.

Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pages 61–73, 1979.

Roger B Myerson. *Game theory*. Harvard university press, 2013.

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*, pages 663–670, 2000.

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

Diederik M Roijers, Shimon Whiteson, and Frans A Oliehoek. Point-based planning for multi-objective pomdps. In *IJCAI 2015: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1666–1672, 2015.

Stuart Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence: a modern approach (Chapter 17.1)*, volume 2. Prentice hall Upper Saddle River, 2003.

Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 101–103. ACM, 1998.

Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

Harold Soh and Yiannis Demiris. Evolving policies for multi-reward partially observable markov decision processes (mr-pomdps). In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 713–720. ACM, 2011.

Gwo-Hshiung Tzeng and Jih-Jeng Huang. *Multiple attribute decision making: methods and applications*. CRC press, 2011.

Weijia Wang. *Multi-objective sequential decision making*. PhD thesis, Université Paris Sud-Paris XI, 2014.

Kyle Hollins Wray and Shlomo Zilberstein. Multi-objective pomdps with lexicographic reward preferences. In *Proceedings of the 24th International Joint Conference of Artificial Intelligence (IJCAI)*, pages 1719–1725, 2015.

Chongjie Zhang and Julie A Shah. Fairness in multi-agent sequential decision-making. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2014.