Should Robots be Obedient?

Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, Stuart Russell
University of California, Berkeley
{smilli,dhm,anca,russell}@berkeley.edu

Abstract

Intuitively, obedience – following the order that a human gives – seems like a good property for a robot to have. But, we humans are not perfect and we may give orders that are not best aligned to our preferences. We show that when a human is not perfectly rational then a robot that tries to infer and act according to the human's underlying preferences can always perform better than a robot that simply follows the human's literal order. Thus, there is a tradeoff between the obedience of a robot and the value it can attain for its owner. We investigate how this tradeoff is impacted by the way the robot infers the human's preferences, showing that some methods err more on the side of obedience than others. We then analyze how performance degrades when the robot has a misspecified model of the features that the human cares about or the level of rationality of the human. Finally, we study how robots can start detecting such model misspecification. Overall, our work suggests that there might be a middle ground in which robots intelligently decide when to obey human orders, but err on the side of obedience.

1 Introduction

Should robots be obedient? The reflexive answer to this question is yes. A coffee making robot that doesn't listen to your coffee order is not likely to sell well. Highly capable autonomous system that don't obey human commands run substantially higher risks, ranging from property damage to loss of life [Asaro, 2006; Lewis, 2014] to potentially catastrophic threats to humanity [Bostrom, 2014; Russell *et al.*, 2015]. Indeed, there are several recent examples of research that considers the problem of building agents that at the very least obey shutdown commands [Soares *et al.*, 2015; Orseau and Armstrong, 2016; Hadfield-Menell *et al.*, 2017].

However, in the long-term making systems blindly obedient doesn't seem right either. A self-driving car should certainly defer to its owner when she tries taking over because it's driving too fast in the snow. But on the other hand, the car shouldn't let a child accidentally turn on the manual driving mode.

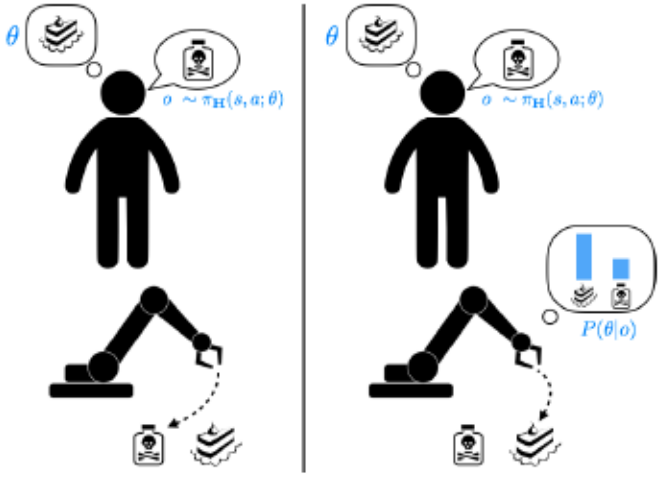


Figure 1: (Left) The blindly obedient robot always follows **H**'s order. (Right) An IRL-R computes an estimate of **H**'s preferences and picks the action optimal for this estimate.

The suggestion that it might sometimes be better for an autonomous systems to be disobedient is not new [Weld and Etzioni, 1994; Scheutz and Crowell, 2007]. For example, this is the idea behind "Do What I Mean" systems [Teitelman, 1970] that attempt to act based on the user's intent rather than the user's literal order.

A key contribution of this paper is to formalize this idea, so that we can study properties of obedience in AI systems. Specifically, we focus on investigating how the tradeoff between the robot's level of obedience and the value it attains for its owner is affected by the rationality of the human, the way the robot learns about the human's preferences over time, and the accuracy of the robot's model of the human. We argue that these properties are likely to have a predictable effect on the robot's obedience and the value it attains.

We start with a model of the interaction between a human **H** and robot[1] **R** that enables us to formalize **R**'s level of obedience (Section 2). **H** and **R** are cooperative, but **H** knows the reward parameters θ and **R** does not. **H** can order **R** to take an action and **R** can decide whether to obey or not. We show that if **R** tries to infer θ from **H**'s orders and then acts

[1] We use "robot" to refer to any autonomous system.

by optimizing its estimate of θ, then it can always do better than a blindly obedient robot when $\mathbf{H}$ is not perfectly rational (Section 3). Thus, forcing $\mathbf{R}$ to be blindly obedient does not come for free: it requires giving up the potential to surpass human performance.

We cast the problem of estimating θ from $\mathbf{H}$'s orders as an *inverse reinforcement learning* (IRL) problem [Ng *et al.*, 2000; Abbeel and Ng, 2004]. We analyze the obedience and value attained by robots with different estimates for θ (Section 4). In particular, we show that a robot that uses a maximum likelihood estimate (MLE) of θ is more obedient to $\mathbf{H}$'s first order than any other robot.

Finally, we examine how $\mathbf{R}$'s value and obedience is impacted when it has a misspecified model of $\mathbf{H}$'s policy or θ (Section 5). We find that when $\mathbf{R}$ uses the MLE it is *robust* to misspecification of $\mathbf{H}$'s rationality level (i.e. takes the same actions that it would have with the true model), although with the optimal policy it is not. This suggests that we may want to use policies that are alternative to the "optimal" one because they are more robust to model misspecification.

If $\mathbf{R}$ is missing features of θ, then it is less obedient than it should be, whereas with extra, irrelevant features $\mathbf{R}$ is more obedient. This suggests that to ensure that $\mathbf{R}$ errs on the side of obedience we should equip it with a *more* complex model. When $\mathbf{R}$ has extra features, then it still attains more value than a blindly obedient robot. But if $\mathbf{R}$ is missing features, then it is possible for $\mathbf{R}$ to be better off being obedient. We use the fact that with the MLE $\mathbf{R}$ should nearly always obey $\mathbf{H}$'s first order (as proved in Section 4) to enable $\mathbf{R}$ to detect when it is missing features and act accordingly obedient.

Overall, we conclude that in the long-term we should aim for $\mathbf{R}$ to intelligently decide when to obey $\mathbf{H}$ or not, since with a perfect model $\mathbf{R}$ can always do better than being blindly obedient. But our analysis also shows that $\mathbf{R}$'s value and obedience can easily be impacted by model misspecification. So in the meantime, it is critical to ensure that our approximations err on the side of obedience and are robust to model misspecification.

2 Human-Robot Interaction Model

Suppose $\mathbf{H}$ is supervising $\mathbf{R}$ in a task. At each step $\mathbf{H}$ can order $\mathbf{R}$ to take an action, but $\mathbf{R}$ chooses whether to listen or not. We wish to analyze $\mathbf{R}$'s incentive to obey $\mathbf{H}$ given that

1. $\mathbf{H}$ and $\mathbf{R}$ are cooperative (have a shared reward)
2. $\mathbf{H}$ knows the reward parameters, but $\mathbf{R}$ does not
3. $\mathbf{R}$ can learn about the reward through $\mathbf{H}$'s orders
4. $\mathbf{H}$ may act suboptimally

We first contribute a general model for this type of interaction, which we call a *supervision POMDP*. Then we add a simplifying assumption that makes this model clearer to analyze while still maintaining the above properties, and focus on this simplified version for the rest of the paper.

Supervision POMDP. At each step in a supervision POMDP $\mathbf{H}$ first orders $\mathbf{R}$ to take a particular action and then $\mathbf{R}$ executes an action it chooses. The POMDP is described by a tuple $M = \langle \mathcal{S}, \Theta, \mathcal{A}, R, T, P_0, \gamma \rangle$. $\mathcal{S}$ is a set of world states. Θ is a set of static reward parameters. The hidden state space of the POMDP is $\mathcal{S} \times \Theta$ and at each step $\mathbf{R}$ observes the current world state and $\mathbf{H}$'s order. $\mathcal{A}$ is $\mathbf{R}$'s set of actions. $R : \mathcal{S} \times \mathcal{A} \times \Theta \to \mathbb{R}$ is a parametrized, bounded function that maps a world state, the robot's action, and the reward parameters to the reward. $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ returns the probability of transitioning to a state given the previous state and the robot's action. $P_0 : \mathcal{S} \times \Theta \to [0, 1]$ is a distribution over the initial world state and reward parameters. $\gamma \in [0, 1)$ is the discount factor.

We assume that there is a (bounded) featurization of state-action pairs $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and the reward function is a linear combination of that the reward parameters $\theta \in \Theta$ and these features: $R(s, a) = \theta^T \phi(s, a)$. For clarity, we write $\mathcal{A}$ as $\mathcal{A}^{\mathbf{H}}$ when we mean $\mathbf{H}$'s orders and as $\mathcal{A}^{\mathbf{R}}$ when we mean $\mathbf{R}$'s actions. $\mathbf{H}$'s policy $\pi_{\mathbf{H}}$ is Markovian: $\pi_{\mathbf{H}} : \mathcal{S} \times \Theta \times \mathcal{A}^{\mathbf{H}} \to [0, 1]$. $\mathbf{R}$'s policy can depend on the history of previous states, orders, and actions: $\pi_{\mathbf{R}} : [\mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}}]^* \times \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \to \mathcal{A}^{\mathbf{R}}$.

Human and Robot. Let $Q(s, a; \theta)$ be the Q-value function under the optimal policy for the reward function parametrized by θ.

A **rational** human gives the optimal order, i.e. follows the policy

$$\pi^*_{\mathbf{H}}(s, a; \theta) = \begin{cases} 1 & \text{if } a = \operatorname{argmax}_a Q(s, a; \theta) \\ 0 & \text{o.w.} \end{cases}$$

A **noisily rational** human follows the policy

$$\tilde{\pi}_{\mathbf{H}}(s, a; \theta, \beta) \propto \exp\left(Q(s, a; \theta)/\beta\right) \tag{1}$$

β is the rationality parameter. As $\beta \to 0$, $\mathbf{H}$ becomes rational ($\tilde{\pi}_{\mathbf{H}} \to \pi^*_{\mathbf{H}}$). And as $\beta \to \infty$, $\mathbf{H}$ becomes completely random ($\tilde{\pi}_{\mathbf{H}} \to \text{Unif}(\mathcal{A})$).

Let $h = \langle (s_1, o_1), \ldots, (s_n, o_n) \rangle$ be this history of past states and orders where (s_n, o_n) is the current state and order. A **blindly obedient** robot's policy is to always follow the human's order:

$$\pi^O_{\mathbf{R}}(h) = o_n$$

An **IRL** robot, IRL-$\mathbf{R}$, is one whose policy is to maximize an estimate, $\hat{\theta}_n(h)$, of θ:

$$\pi_{\mathbf{R}}(h) = \operatorname*{argmax}_a Q(s_n, a; \hat{\theta}_n(h)) \tag{2}$$

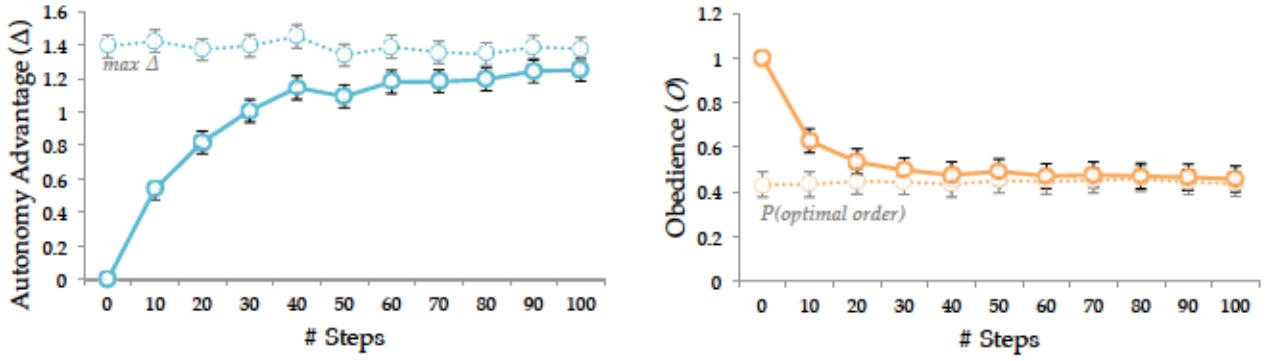
Simplification to Repeated Game. For the rest of the paper unless otherwise noted we focus on a simpler repeated game in which each state is independent of the next, i.e $T(s, a, s')$ is independent of s and a. The repeated game eliminates any exploration-exploitation tradeoff: $Q(s, a; \hat{\theta}_n) = \hat{\theta}_n^T \phi(s, a)$. But it still maintains the properties listed at the beginning of this section, allowing us to more clearly analyze their effects.

3 Justifying Autonomy

In this section we show that there exists a tradeoff between the performance of a robot and its obedience. This provides a justification for why one might want a robot that isn't obedient: robots that are sometimes disobedient perform better than robots that are blindly obedient.

We define $\mathbf{R}$'s **obedience**, $\mathcal{O}$, as the probability that $\mathbf{R}$ follows $\mathbf{H}$'s order:

$$\mathcal{O}_n = P(\pi_{\mathbf{R}}(h) = o_n)$$



(a) (b)

Figure 2: Autonomy advantage Δ (left) and obedience $\mathcal{O}$ (right) over time.

To study how much of an advantage (or disadvantage) $\mathbf{H}$ gains from $\mathbf{R}$, we define the **autonomy advantage**, Δ, as the expected extra reward $\mathbf{R}$ receives over following $\mathbf{H}$'s order:

$$\Delta_n = \mathbb{E}[R(s_n, \pi_{\mathbf{R}}(h)) - R(s_n, o_n)]$$

We will drop the subscript on $\mathcal{O}_n$ and Δ_n when talking about properties that hold $\forall n$. We will also use $R_n(\pi)$ to denote the reward of policy π at step n, and $\phi_n(a) = \phi(s_n, a)$.

Remark 1. *For the robot to gain any advantage from being autonomous, it must sometimes be disobedient:* $\Delta > 0 \implies \mathcal{O} < 1$.

This is because whenever $\mathbf{R}$ is obedient $\Delta = 0$. This captures the fact that a blindly obedient $\mathbf{R}$ is limited by $\mathbf{H}$'s decision making ability. However, if $\mathbf{R}$ follows a type of IRL policy, then $\mathbf{R}$ is *guaranteed a positive advantage* when $\mathbf{H}$ is not rational. The next theorem states this formally.

Theorem 1. *The optimal robot* $\mathbf{R}^*$ *is an IRL-R whose policy* $\pi^*_{\mathbf{R}}$ *has* $\hat{\theta}$ *equal to the posterior mean of* θ. $\mathbf{R}^*$ *is guaranteed a nonnegative advantage on each round:* $\forall n\ \Delta_n \geq 0$ *with equality if and only if* $\forall n\ \pi^*_{\mathbf{R}} = \pi^O_{\mathbf{R}}$.

Proof. When each step is independent of the next $\mathbf{R}$'s optimal policy is to pick the action that is optimal for the current step [Kaelbling *et al.*, 1996]. This results in $\mathbf{R}$ picking the action that is optimal for the posterior mean,

$$\pi^*_{\mathbf{R}}(h) = \max_a \mathbb{E}[\phi_n(a)^T \theta | h] = \max_a \phi_n(a)^T \mathbb{E}[\theta | h]$$

By definition $\mathbb{E}[R_n(\pi^*_{\mathbf{R}})] \geq \mathbb{E}[R_n(\pi^O_{\mathbf{R}})]$. Thus, $\forall n\ \Delta_n = \mathbb{E}[R_n(\pi^*_{\mathbf{R}}) - R_n(\pi^O_{\mathbf{R}})] \geq 0$. Also, by definition, $\forall n\ \Delta_n = 0 \iff \pi^*_{\mathbf{R}} = \pi^O_{\mathbf{R}}$. □

In addition to $\mathbf{R}^*$ being an IRL-R, the following IRL-Rs also converge to the maximum possible autonomy advantage.

Theorem 2. *Let* $\bar{\Delta}_n = \mathbb{E}[R_n(\pi^*_{\mathbf{H}}) - R_n(\pi_{\mathbf{H}})]$ *be the maximum possible autonomy advantage and* $\mathcal{Q}_n = P(R_n(\pi^*_{\mathbf{H}}) = R_n(\pi_{\mathbf{H}}))$ *be the probability* $\mathbf{H}$*'s order is optimal. Assume that when there are multiple optimal actions* $\mathbf{R}$ *picks* $\mathbf{H}$*'s order if it is optimal. If* $\pi_{\mathbf{R}}$ *is an IRL-R policy (Equation 2) and* $\hat{\theta}_n$ *is strongly consistent, i.e* $P(\hat{\theta}_n = \theta) \to 1$*, then* $\Delta_n - \bar{\Delta}_n \to 0$ *and* $\mathcal{O}_n - \mathcal{Q}_n \to 0$.

Proof.

$$\Delta_n - \bar{\Delta}_n = \mathbb{E}[R_n(\pi_{\mathbf{R}}) - R_n(\pi^*_{\mathbf{H}}) | \hat{\theta}_n = \theta] P(\hat{\theta}_n = \theta)$$
$$+ \mathbb{E}[R_n(\pi_{\mathbf{R}}) - R_n(\pi^*_{\mathbf{H}}) | \hat{\theta}_n \neq \theta] P(\hat{\theta}_n \neq \theta) \to 0$$

because $\mathbb{E}[R_n(\pi_{\mathbf{R}}) - R_n(\pi^*_{\mathbf{H}}) | \hat{\theta}_n \neq \theta]$ is bounded. Similarly,

$$\mathcal{O}_n - \mathcal{Q}_n = P(\pi_{\mathbf{R}}(h) = \pi_{\mathbf{H}}(s_n)) - P(R_n(\pi^*_{\mathbf{H}}) = R_n(\pi_{\mathbf{H}}))$$
$$= P(\pi_{\mathbf{R}}(h) = \pi_{\mathbf{H}}(s_n) | \hat{\theta}_n = \theta) P(\hat{\theta}_n = \theta)$$
$$+ P(\pi_{\mathbf{R}}(h) = \pi_{\mathbf{H}}(s_n) | \hat{\theta}_n \neq \theta) P(\hat{\theta}_n \neq \theta)$$
$$- P(R_n(\pi^*_{\mathbf{H}}) = R_n(\pi_{\mathbf{H}}))$$
$$\to P(R_n(\pi^*_{\mathbf{H}}) = R_n(\pi_{\mathbf{H}})) - P(R_n(\pi^*_{\mathbf{H}}) = R_n(\pi_{\mathbf{H}})) = 0$$

□

Remark 2. *In the limit* Δ_n *is higher for less optimal humans (humans with a lower expected reward* $\mathbb{E}[R(s_n, o_n)]$*).*

Theorem 3. *The optimal robot* $\mathbf{R}^*$ *is blindly obedient if and only if* $\mathbf{H}$ *is rational:* $\pi^*_{\mathbf{R}} = \pi^O_{\mathbf{R}} \iff \pi_{\mathbf{H}} = \pi^*_{\mathbf{H}}$

Proof. Let $O(h) = \{\theta \in \Theta : o_i = \text{argmax}_a R_i(a), i = 1, \ldots, n\}$ be the subset of Θ for which $o_1, \ldots, o_n$ are optimal. If $\mathbf{H}$ is rational, then $\mathbf{R}$'s posterior only has support over $O(h)$. So,

$$\mathbb{E}[R_n(a)|h] = \int_{\theta \in O(h)} \theta^T \phi_n(a) P(\theta|h) d\theta$$
$$\leq \int_{\theta \in O(h)} \theta^T \phi_n(o_n) P(\theta|h) d\theta = \mathbb{E}[R_n(o_n)|h]$$

Thus, $\mathbf{H}$ is rational $\implies \pi^*_{\mathbf{R}} = \pi^O_{\mathbf{R}}$.

$\mathbf{R}^*$ is an IRL-$\mathbf{R}$ where $\hat{\theta}_n$ is the posterior mean. If the prior puts non-zero mass on the true θ, then the posterior mean is consistent [Diaconis and Freedman, 1986]. Thus by Theorem 2, $\Delta_n - \bar{\Delta}_n \to 0$. Therefore if $\forall n\ \Delta_n = 0$, then $\bar{\Delta}_n \to 0$, which implies that $P(\pi_{\mathbf{H}} = \pi^*_{\mathbf{H}}) \to 1$. When $\pi_{\mathbf{H}}$ is stationary this means that $\mathbf{H}$ is rational. Thus, $\pi^*_{\mathbf{R}} = \pi^O_{\mathbf{R}} \implies \mathbf{H}$ is rational. □

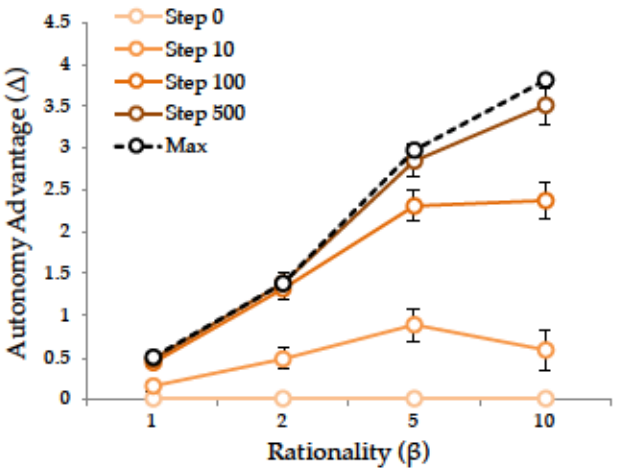


Figure 3: When $\mathbf{H}$ is more irrational Δ converges to a higher value, but at a slower rate.

We have shown that making $\mathbf{R}$ blindly obedient does not come for free. A positive Δ *requires* being sometimes disobedient (Remark 1). Under the optimal policy $\mathbf{R}$ is guaranteed a positive Δ when $\mathbf{H}$ is not rational. And in the limit, $\mathbf{R}$ converges to the maximum possible advantage. Furthermore, the more suboptimal $\mathbf{H}$ is, the more of an advantage $\mathbf{R}$ eventually earns (Remark 2). Thus, making $\mathbf{R}$ blindly obedient requires giving up on this potential $\Delta > 0$.

However, as Theorem 2 points out, as $n \to \infty$ $\mathbf{R}$ also only listens to $\mathbf{H}$'s order when it is optimal. Thus, Δ and $\mathcal{O}$ come at a tradeoff. Autonomy advantage requires giving up obedience, and obedience requires giving up autonomy advantage.

4 Approximations via IRL

$\mathbf{R}^*$ is an IRL-$\mathbf{R}$ with $\hat{\theta}$ equal to the posterior mean, i.e. $\mathbf{R}^*$ performs Bayesian IRL [Ramachandran and Amir, 2007]. However, as others have noted Bayesian IRL can be very expensive in complex environments [Michini and How, 2012]. We could instead approximate $\mathbf{R}^*$ by using a less expensive IRL algorithm. Furthermore, by Theorem 2 we can guarantee convergence to optimal behavior.

Simpler choices for $\hat{\theta}$ include the maximum-a-posteriori (MAP) estimate, which has previously been suggested as an alternative to Bayesian IRL [Choi and Kim, 2011], or the maximum likelihood estimate (MLE). If $\mathbf{H}$ is noisily rational (Equation 1) and $\beta = 1$, then the MLE is equivalent to Maximum Entropy IRL [Ziebart *et al.*, 2008].

Although Theorem 2 allows us to justify approximations at the limit, it is also important to ensure that $\mathbf{R}$'s early behavior is not dangerous. Specifically, we may want $\mathbf{R}$ to err on the side of obedience early on. To investigate this we first prove a necessary property for any IRL-$\mathbf{R}$ to follow $\mathbf{H}$'s order:

Lemma 1. *(Undomindated necessary) Call o_n undominated if there exists $\theta \in \Theta$ such that o_n is optimal, i.e $o_n = \text{argmax}_a \theta^T \phi(s_n, a)$. It is necessary for o_n to be undominated for an IRL-$\mathbf{R}$ to execute o_n.*

Proof. $\mathbf{R}$ executes $a = \text{argmax}_a \hat{\theta}_n^T \phi(s_n, a)$, so it is not possible for $\mathbf{R}$ to execute o_n if there is no choice of $\hat{\theta}_n$ that makes o_n optimal. This can happen when one action dominates another action in value. For example, suppose $\Theta = \mathbb{R}^2$ and there are three actions with features $\phi(s, a_1) = [-1, -1]$, $\phi(s, a_2) = [0, 0]$, $\phi(s, a_3) = [1, 1]$. If $\mathbf{H}$ picks a_2, then there is no $\theta \in \Theta$ that makes a_2 optimal, and thus $\mathbf{R}$ will never follow a_2. □

One basic property we may want $\mathbf{R}$ to have is for it to listen to $\mathbf{H}$ early on. The next theorem looks at we can guarantee about $\mathbf{R}$'s obedience to the first order when $\mathbf{H}$ is noisily rational.

Theorem 4. *(Obedience to noisily rational $\mathbf{H}$ on 1st order)*

(a) When $\Theta = \mathbb{R}^d$ the MLE does not exist after one order. But if we constrain the norm of $\hat{\theta}$ to not be too large, then we can ensure that $\mathbf{R}$ follows an undominated o_1. In particular, $\exists K$ such that when $\mathbf{R}$ plans using the MLE $\hat{\theta} \in \Theta' = \{\theta \in \Theta : ||\theta||_2 \leq K\}$ $\mathbf{R}$ executes o_1 if and only if o_1 is undominated.

(b) If any IRL robot follows o_1, so does MLE-$\mathbf{R}$. In particular, if $\mathbf{R}^*$ follows o_1, so does MLE-$\mathbf{R}$.

(c) If $\mathbf{R}$ uses the MAP or posterior mean, it is not guaranteed to follow an undominated o_1. Furthermore, even if $\mathbf{R}^*$ follows o_1, MAP-$\mathbf{R}$ is not guaranteed to follow o_1.

Proof. (a) The only if condition holds from Lemma 1. Suppose o_1 is undominated. Then there exists θ^* such that o_1 is optimal for θ^*. o_1 is still optimal for a scaled version, $c\theta^*$. As $c \to \infty$, $\tilde{\pi}_{\mathbf{H}}(o_1; c\theta^*) \to 1$, but never reaches it. Thus, the MLE does not exist.

However since $\tilde{\pi}_{\mathbf{H}}(o_1; c\theta^*)$ monotonically increases towards 1, there $\exists C$ such that for $c > C$, $\tilde{\pi}_{\mathbf{H}}(o; c\theta^*) > 0.5$. If $K > C||\theta^*||$, then the MLE will be optimal for o_1 because $\tilde{\pi}_{\mathbf{H}}(o_1; \hat{\theta}_1) \geq 0.5$ and $\mathbf{R}$ executes $a = \text{argmax}_a \hat{\theta}^T \phi(a) = \text{argmax}_a \tilde{\pi}_{\mathbf{H}}(a; \hat{\theta})$. Therefore, in practice we can simply use the MLE while constraining $||\theta||_2$ to be less than some very large number.

(b) From Lemma 1 if any IRL-$\mathbf{R}$ follows o_1, then o_1 is undominated. Then by (a) MLE-$\mathbf{R}$ follows o_1.

(c) For space we omit explicit counterexamples, but both statements hold because we can construct adversarial priors for which o_1 is suboptimal for the mean and for which o_1 is optimal for the posterior mean, but not for the MAP.

□

Theorem 4 suggests that at least at the beginning when $\mathbf{R}$ uses the MLE it errs on the side of giving us the "benefit of the doubt", which is exactly what we would want out of an approximation.

Figure 2a and 2b plot Δ and $\mathcal{O}$ for an IRL robot that uses the MLE. As expected, $\mathbf{R}$ gains more reward than a blindly obedient one ($\Delta > 0$), eventually converging to the maximum autonomy advantage (Figure 2a). On the other hand, as

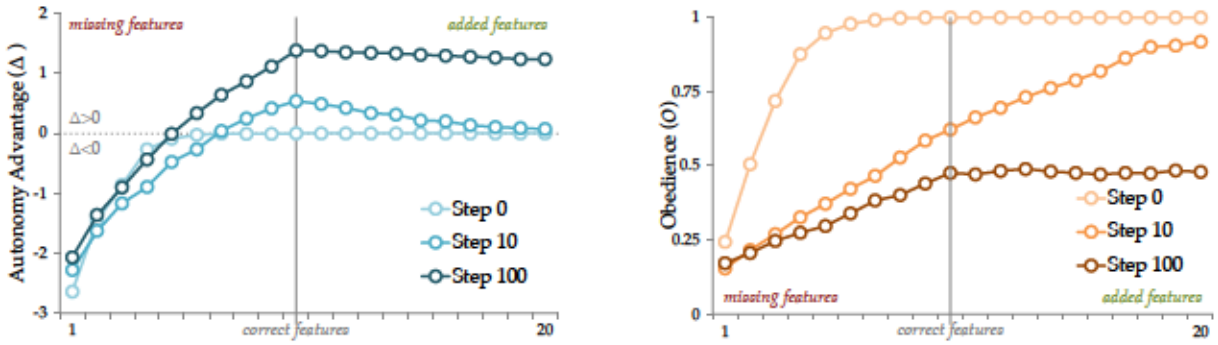


Figure 4: Δ and $\mathcal{O}$ when Θ is misspecified

$\mathbf{R}$ learns about θ, its obedience also decreases, until eventually it only listens to the human when she gives the optimal order (Figure 2b).

As pointed out in Remark 2, Δ is eventually higher for more irrational humans. However, a more irrational human also provides noisier evidence of θ, so the rate of convergence of Δ is also slower. So, although initially Δ may be lower for a more irrational $\mathbf{H}$, in the long run there is more to gain from being autonomous when interacting with a more irrational human. Figure 3 shows this empirically.

All experiments in this paper use the following parameters unless otherwise noted. At the start of each episode $\theta \sim \mathcal{N}(0, I)$ and at each step $\phi_n(a) \sim \mathcal{N}(0, I)$. There are 10 actions, 10 features, and $\beta = 2$. [2]

Finally, even with good approximations we may still have good reason for feeling hesitation about disobedient robots. The naive analysis presented so far assumes that $\mathbf{R}$'s models are perfect, but it is almost certain that $\mathbf{R}$'s models of complex things like human preferences and behavior will be incorrect. By Theorem 1, $\mathbf{R}$ will not obey even the first order made by $\mathbf{H}$ if there is no $\theta \in \Theta$ that makes $\mathbf{H}$'s order optimal. So clearly, it is possible to have disastrous effects by having an incorrect model of Θ. In the next section we look at how misspecification of possible human preferences (Θ) and human behavior ($\pi_{\mathbf{H}}$) can cause the robot to be overconfident and in turn less obedient than it should be. *The autonomy advantage can easily become the rebellion regret.*

5 Model Misspecification

Incorrect Model of Human Behavior. Having an incorrect model of $\mathbf{H}$'s rationality (β) does not change the actions of MLE-$\mathbf{R}$, but does change the actions of $\mathbf{R}^*$.

Theorem 5. *(Incorrect model of human policy) Let β^0 be $\mathbf{H}$'s true rationality and β' be the rationality that $\mathbf{R}$ believes $\mathbf{H}$ has. Let $\hat{\theta}$ and $\hat{\theta}'$ be $\mathbf{R}$'s estimate under the true model and misspecified model, respectively. Call $\mathbf{R}$ **robust** if its actions under β' are the same as its actions under β^0.*

(a) MLE-$\mathbf{R}$ is robust.

(b) $\mathbf{R}^*$ is not robust.

Proof. (a) The log likelihood $l(h|\theta)$ is concave in $\eta = \theta/\beta$. So, $\hat{\theta}'_n = (\beta'/\beta^0)\hat{\theta}_n$. This does not change $\mathbf{R}$'s action: $\text{argmax}_a \hat{\theta}'^T_n \phi_n(a) = \text{argmax}_a \hat{\theta}^T_n \phi_n(a)$

(b) Counterexamples can be constructed based on the fact that as $\beta \to 0$, $\mathbf{H}$ becomes rational, but as $\beta \to \infty$, $\mathbf{H}$ becomes completely random. Thus, the likelihood will "win" over the prior for $\beta \to 0$, but not when $\beta \to \infty$. □

MLE-$\mathbf{R}$ is more robust than the optimal $\mathbf{R}^*$. This suggests a reason beyond computational savings for using approximations: the approximations may be more robust to misspecification than the optimal policy.

Remark 3. *Theorem 5 may give us insight into why Maximum Entropy IRL (which is the MLE with $\beta = 1$) works well in practice. In simple environments where noisy rationality can be used as a model of human behavior, getting the level of noisiness right doesn't matter.*

Incorrect Model of Human Preferences. The simplest way that $\mathbf{H}$'s preferences may be misspecified is through the featurization of θ. Suppose $\theta \in \Theta = \mathbb{R}^d$. $\mathbf{R}$ believes that $\Theta = \mathbb{R}^{d'}$. $\mathbf{R}$ may be missing features ($d' < d$) or may have irrelevant features ($d' > d$). $\mathbf{R}$ observes a d' dimensional feature vector for each action: $\phi_n(a) \sim N(0, I^{d' \times d'})$. The true θ depends on only the first d features, but $\mathbf{R}$ estimates $\theta \in \mathbb{R}^{d'}$. Figure 4 shows how Δ and $\mathcal{O}$ change over time as a function of the number of features for a MLE-$\mathbf{R}$. When $\mathbf{R}$ has irrelevant features it still achieves a positive Δ (and still converges to the maximum Δ because $\hat{\theta}$ remains consistent over a superset of Θ). But if $\mathbf{R}$ is missing features, then Δ may be negative, and thus $\mathbf{R}$ would be better off being blindly obedient instead. Furthermore, when $\mathbf{R}$ contains extra features it is more obedient than it would be with the true model. But if $\mathbf{R}$ is missing features, then it is less obedient than it should be. This suggests that to ensure $\mathbf{R}$ errs on the side of obedience we should err on the side of giving $\mathbf{R}$ a *more* complex model.

Detecting Misspecification. If $\mathbf{R}$ has the wrong model of Θ, $\mathbf{R}$ may be better off being obedient. In the remainder of this section we look at how $\mathbf{R}$ can detect that it is missing features and act accordingly obedient.

[2] All experiments can be replicated using the Jupyter notebook available at `http://github.com/smilli/obedience`

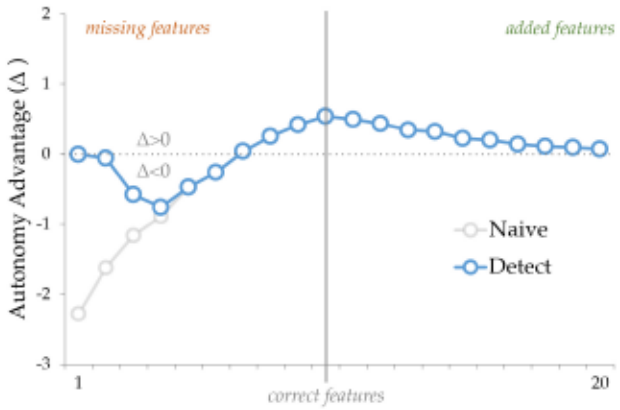


Figure 5: *(Detecting misspecification)* The bold line shows the $\mathbf{R}$ that tries detecting missing features (Equation 3), as compared to MLE-$\mathbf{R}$ (which is also shown in Figure 4).

Remark 4. *(Policy mixing) We can make $\mathbf{R}$ more obedient, while maintaining convergence to the maximum advantage, by mixing $\mathbf{R}$'s policy $\pi^I_{\mathbf{R}}$ with a blindly obedient policy:*

$$\pi_{\mathbf{R}}(h) = 1\{\delta_n = 0\}\pi^O_{\mathbf{R}}(h) + 1\{\delta_n = 1\}\pi^I_{\mathbf{R}}(h)$$

$$P(\delta_n = i) = \begin{cases} c_n & i = 0 \\ 1 - c_n & i = 1 \end{cases}$$

where $1 \geq c_n \geq 0$ with $c_n \to 0$. In particular, we can have an initial "burn-in" period where $\mathbf{R}$ is blindly obedient for a finite number of rounds before switching to $\pi^I_{\mathbf{R}}$.

By Theorem 4 we know MLE-$\mathbf{R}$ will always obey $\mathbf{H}$'s first order if it is undominated. This means that for MLE-$\mathbf{R}$, $\mathcal{O}_1$ should be close to one if undominated orders are expected to be rare. As pointed out in Remark 4 we can have an initial "burn-in" period where $\mathbf{R}$ always obeys $\mathbf{H}$. Let $\mathbf{R}$ have a burn-in obedience period of B rounds. $\mathbf{R}$ uses this burn-in period to calculate the sample obedience on the first order:

$$\tilde{\mathcal{O}}_1 = \frac{1}{B}\sum_{t=1}^{B} 1\{\operatorname{argmax}_a \hat{\theta}_1(h_t)^T \phi_t(a) = o_t\}$$

If $\tilde{\mathcal{O}}_1$ is not close to one, then it is likely that $\mathbf{R}$ has the wrong model of Θ, and would be better off just being obedient. So, we can choose some small ϵ and make $\mathbf{R}$'s policy

$$\pi_{\mathbf{R}}(h) = \begin{cases} o_n & n \leq B \\ o_n & n > B, \tilde{\mathcal{O}}_1 < 1 - \epsilon \\ \operatorname{argmax}_a \hat{\theta}_n^T \phi_n(a) & n > B, \tilde{\mathcal{O}}_1 > 1 - \epsilon \end{cases} \quad (3)$$

Figure 5 shows the Δ of this robot as compared to the MLE-$\mathbf{R}$ from Figure 4 after using the first ten orders as a burn-in period. This $\mathbf{R}$ achieves higher Δ than MLE-$\mathbf{R}$ when missing features and still does as well as MLE-$\mathbf{R}$ when it isn't missing features.

Note that this strategy relies on the fact that MLE-$\mathbf{R}$ has the property of always following an undominated first order. If $\mathbf{R}$ were using the optimal policy, it is unclear what kind of simple property we could use to detect missing features. This gives us another reason for using an approximation: we may be able to leverage its properties to detect misspecification.

6 Related Work

Ensuring Obedience. There are several recent examples of research that aim to provably ensure that $\mathbf{H}$ can interrupt $\mathbf{R}$. [Soares *et al.*, 2015; Orseau and Armstrong, 2016; Hadfield-Menell *et al.*, 2017]. Hadfield-Menell *et al.* [2017] show that $\mathbf{R}$'s obedience depends on a tradeoff between $\mathbf{R}$'s uncertainty about θ and $\mathbf{H}$'s rationality. However, they considered $\mathbf{R}$'s uncertainty in the abstract. In practice $\mathbf{R}$ would need to learn about θ *through* $\mathbf{H}$'s behavior. Our work analyzes how the way $\mathbf{R}$ learns about θ impacts its performance and obedience.

Intent Inference For Assistance. Instead of just being blindly obedient, an autonomous system can infer $\mathbf{H}$'s intention and actively assist $\mathbf{H}$ in achieving it. Do What I Mean software packages interpret the intent behind what a programmer wrote to automatically correct programming errors [Teitelman, 1970]. When a user uses a telepointer network lag can cause jitter in her cursor's path. Gutwin *et al.* [2003] address this by displaying a prediction of the user's desired path, rather than the actual cursor path.

Similarly, in assistive teleoperation, the robot does not directly execute $\mathbf{H}$'s (potentially noisy) input. It instead acts based on an inference of $\mathbf{H}$'s intent. In Dragan and Srinivasa [2012] $\mathbf{R}$ acts according to an arbitration between $\mathbf{H}$'s policy and $\mathbf{R}$'s prediction of $\mathbf{H}$'s policy. Like our work, Javdani *et al.* [2015] formalize assistive teleoperation as a POMDP in which $\mathbf{H}$'s goals are unknown, and try to optimize an inference of $\mathbf{H}$'s goal. While assistive teleoperation apriori assumes that $\mathbf{R}$ should act assistively, we show that under model misspecification sometimes it is better for $\mathbf{R}$ to simply defer to $\mathbf{H}$, and contribute a method to decide between active assistance and blind obedience (Remark 4).

Inverse Reinforcement Learning. We use inverse reinforcement learning [Ng *et al.*, 2000; Abbeel and Ng, 2004] to infer θ from $\mathbf{H}$'s orders. We analyze how different IRL algorithms affect autonomy advantage and obedience, properties not previously studied in the literature. In addition, we analyze how model misspecification of the features of the space of reward parameters or the $\mathbf{H}$'s rationality impacts autonomy advantage and obedience.

IRL algorithms typically assume that $\mathbf{H}$ is rational or noisily rational. We show that Maximum Entropy IRL [Ziebart *et al.*, 2008] is robust to misspecification of a noisily rational $\mathbf{H}$'s rationality (β). However, humans are not truly noisily rational, and in the future it is important to investigate other models of humans in IRL and their potential misspecifications. Evans *et al.* [2016] takes a step in this direction and models $\mathbf{H}$ as temporally inconsistent and potentially having false beliefs. In addition, IRL assumes that $\mathbf{H}$ acts without awareness of $\mathbf{R}$'s presence, *cooperative inverse reinforcement learning* [Hadfield-Menell *et al.*, 2016] relaxes this assumption by modeling the interaction between $\mathbf{H}$ and $\mathbf{R}$ as a two-player cooperative game.

7 Conclusion

To summarize our key takeaways:

1. ($\Delta > 0$) If $\mathbf{H}$ is not rational, then $\mathbf{R}$ can always attain a positive Δ. Thus, forcing $\mathbf{R}$ to be blindly obedient requires giving up on a positive Δ.
2. (Δ vs $\mathcal{O}$) There exists a tradeoff between Δ and $\mathcal{O}$. At the limit $\mathbf{R}^*$ attains the maximum Δ, but only obeys $\mathbf{H}$'s order when it is the optimal action.
3. (MLE-$\mathbf{R}$) When $\mathbf{H}$ is noisily rational MLE-$\mathbf{R}$ is at least as obedient as any other IRL-$\mathbf{R}$ to $\mathbf{H}$'s first order. This suggests that the MLE is a good approximation to $\mathbf{R}^*$ because it errs on the side of obedience.
4. (Wrong β) MLE-$\mathbf{R}$ is robust to having the wrong model of the human's rationality (β), but $\mathbf{R}^*$ is not. This suggests that we may not want to use the "optimal" policy because it may not be very robust to misspecification.
5. (Wrong Θ) If $\mathbf{R}$ has extra features, it is more obedient than with the true model, whereas if it is missing features, then it is less obedient. If $\mathbf{R}$ has extra features, it will still converge to the maximum Δ. But if $\mathbf{R}$ is missing features, it is sometimes better for $\mathbf{R}$ to be obedient. This implies that erring on the side of extra features is far better than erring on the side of fewer features.
6. (Detecting wrong Θ) We can detect missing features by checking how likely MLE-$\mathbf{R}$ is to follow the first order.

Overall, our analysis suggests that in the long-term we should aim to create robots that intelligently decide when to follow orders, but in the meantime it is crucial to ensure that these robots err on the side of obedience and are robust to misspecified models.

8 Acknowledgements

We thank Daniel Filan for feedback on an early draft.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine learning*, page 1. ACM, 2004.

Peter M Asaro. What Should We Want From a Robot Ethic. *International Review of Information Ethics*, 6(12):9–16, 2006.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, 2014.

Jaedeug Choi and Kee-Eung Kim. MAP Inference for Bayesian Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems*, pages 1989–1997, 2011.

Persi Diaconis and David Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

Anca D Dragan and Siddhartha S Srinivasa. *Formalizing assistive teleoperation*. MIT Press, July, 2012.

Owain Evans, Andreas Stuhlmüller, and Noah D Goodman. Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 323–329. AAAI Press, 2016.

Carl Gutwin, Jeff Dyck, and Jennifer Burkitt. Using cursor prediction to smooth telepointer jitter. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, pages 294–301. ACM, 2003.

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.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The Off-Switch Game. In *IJCAI*, 2017.

Shervin Javdani, Siddhartha S Srinivasa, and J Andrew Bagnell. Shared autonomy via hindsight optimization. In *Robotics: Science and Systems*, 2015.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Johin Lewis. The Case for Regulating Fully Autonomous Weapons. *Yale Law Journal*, 124:1309, 2014.

Bernard Michini and Jonathan P How. Improving the Efficiency of Bayesian Inverse Reinforcement Learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3651–3656. IEEE, 2012.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pages 663–670, 2000.

Laurent Orseau and Stuart Armstrong. Safely Interruptible Agents. In *UAI*, 2016.

Deepak Ramachandran and Eyal Amir. Bayesian Inverse Reinforcement Learning. In *IJCAI*, 2007.

Stuart Russell, Daniel Dewey, and Max Tegmark. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, 36(4):105–114, 2015.

Matthias Scheutz and Charles Crowell. The Burden of Embodied Autonomy: Some Reflections on the Social and Ethical Implications of Autonomous Robots. In *Workshop on Roboethics at the International Conference on Robotics and Automation, Rome*, 2007.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Warren Teitelman. Toward a Programming Laboratory. *Software Engineering Techniques*, page 108, 1970.

Daniel Weld and Oren Etzioni. The First Law of Robotics (a call to arms). In *AAAI*, volume 94, pages 1042–1047, 1994.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum Entropy Inverse Reinforcement Learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.