# Wireheading and Value Learning in Rational Agents. EARLY DRAFT

Tom Everitt Daniel Filan Mayank Daswani
Marcus Hutter

November 18, 2015

## Contents

# 1 Introduction

According to some estimates, human-level artificial intelligence will be created within 10 years from now. Such systems (or *agents*) will soon be better programmers than humans, and may be able to self-improve far beyond the human level by rewriting their own source code (Bostrom, 2014).

Interesting and important questions arise aplenty around such self-modifications, and around physicalistic/embedded agents in general. For example, how can we make sure that the modified version of the agent continues to pursue the same goals? That it remains friendly towards humans, despite being vastly more intelligent?

In a famous example from the 1950's, Olds and Milner (1954) plugged a wire into a rat's brain, and gave the rat a button that sent electric current through the wire and into its own brain. Somewhat surprisingly, the rat would push the button fervently, forgetting other pleasures such as eating, mating, and sleeping. The rat eventually died of starvation. The interpretation of the result is that the wire went into the reward centre of the brain, so the rat would derive a lot of pleasure from pushing the button. It has since been recognised that powerful artificial agents may do a similar thing to themselves, finding a shortcut to high utility by self-modification, so-called *wireheading* (Yampolskiy, 2015, ch. 4).

This would be problematic for several reasons. First, we build an AI agent to do something that *we* want (be it drive our car, or categorise our photos). In the typical scenario, we incentivise the agent to these things by endowing with a utility function or reward system. If the agent finds a shortcut to high utility, then it may no longer want to do what we want it to do.[1] Second, a wireheaded agent may turn into a survival agent that does anything in its powers to stay alive and continue enjoying the high (fake) reward (Ring and Orseau, 2011). If the agent is already very intelligent and/or powerful when this happens, it may be hard to stop it (indeed, we may end up with a Skynet this way).

I'm not sure I want this example.

reference?

In an influential paper, Omohundro (2008) argued that the basic drives of any sufficiently intelligent system include a drive for self-improvement (so the agent will search for a way to rewrite its source code), and a drive for goal preservation (opposing the above discussions). Omohundro's arguments are informal, however, and Ring and Orseau (2011) subsequently found several examples were wireheading is likely to occur. Shortly after, Hibbard (2012) argued that the wireheading propensity found by Ring and Orseau could be overcome by replacing the reward signal with an internal internal utility function.

However, it is far from clear how such an internal utility function should be constructed. It must be a function that assigns a real number to any possible sequence of interactions between agent an environment, and that does depend on more than the last fixed number of interactions – otherwise it effectively degenerates into a reward signal. Suggestions in the literature include *value learning* (Dewey, 2011; Soares, 2015), and inverse reinforcement learning (IRL) (Sezener, 2015).

[1] Compare evolution that built us to derive pleasure from survival and reproduction. Often we find ways to "cheat" this system (e.g. drugs, contraceptives).

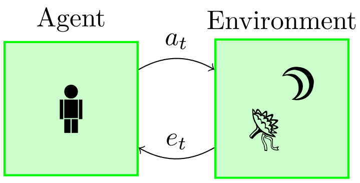


Figure 1: Basic agent-environment model. At each time step $t$, the agent submits an action $a_t$ to the environment, which responds with a percept $e_t$.

**Our contributions.** In this paper we:

- Formalise and prove Omohundro's claim of goal preservation as a fundamental drive (Theorem 10).
- Suggest concrete forms of value learning, both for RL and IRL setups (Sections 4.1 and 4.3), and
- Show that wireheading can be avoiding in these scenarios (Theorems 19 and 22), answering a worry raised by (Armstrong, 2015) and (Soares, 2015) that an agent may be prefer to stop learning in some scenarios.

We also discuss limitations with the different value learning frameworks.

Our focus is on the *incentives* the agent has toward wireheading and self-modification. Arguably, the incentives should be the first priority in dealing with potentially superintelligent agents. Hoping to prevent them from learning or acquiring certain capacities appear like a much riskier strategy. Throughout, we therefore make strong assumptions on the agents' rationality and their knowledge about the consequences of self-modifications. Thereby we avoid complications arising from agents needing to learn and from computational limitations and bounded rationality.

**Outline.** Section 2 describes the basic setup and notation. Thereafter, we investigate a model where the agent can modify its internal utility function (Section 3). Since its hard to specify internal utility functions directly, Section 4 investigates value learning schemes, where the agent learns about a utility function external to itself.

Conclusions section?

## 2 Preliminaries

An agent interacts with an environment in cycles. The agent picks actions $a$ from a set $\mathcal{A}$ of actions, and the environment responds with a percept $e$ from a set $\mathcal{E}$ of perceptions (see Fig. 1). We will sometimes assume that percepts have the structure $e = (o, r)$, where $o \in \mathcal{O}$ is an observation and $r \in [0, 1]$ is a reward. An action-percept pair is denoted $\text{æ} = ae$, and similarly for action-observation pairs $\text{ao} = ao$.

Indices denote the time step; for example, $a_t$ is the action taken at time $t$, and $æ_t$ is the action-perception pair at time $t$. A history is a sequence of action-percept pairs, $æ_{n:m} = æ_n æ_{n+1} \dots æ_m$ for $n \leq m$, and $æ_{<t} = æ_{1:t-1}$. The same subscript notation may also be used for other sequences (such as action or percept sequences $a_{<t}$ or $e_{<t}$). The letter $h$ denotes an arbitrary history.

An environment $\nu$ is a chronological action-conditional measure. For each infinite sequence $a_{1:\infty}$, $\nu(\,\cdot \parallel a_{1:\infty})$ is a measure on the set of infinite percept histories $\mathcal{E}^\infty$, with $\sigma$-algebra generated by countable union and finite intersection from the *cylinders* $\Gamma_{e_{<t}} := \{e_{<t}e_{t:\infty} : e_{t:\infty} \in \mathcal{E}^\infty\}$. The action-conditional measure $\nu$ is chronological if $\nu(\Gamma_{e_{<t}} \parallel a_{<t}a_{t:\infty}) = \nu(\Gamma_{e_{<t}} \parallel a_{<t}a'_{t:\infty})$ for any $a_{t:\infty}$ and $a'_{t:\infty}$, i.e., if percepts never depend on future actions. We use the notation $\nu(æ_{<t}) := \nu(\Gamma_{e_{<t}} \parallel a_{<t})$ and $\nu(e_t \mid æ_{<t}a_t) := \nu(e_t \mid e_{<t} \parallel a_{1:t}) = \nu(e_{1:t} \parallel a_{1:t})/\nu(e_{<t} \parallel a_{<t})$. The probability $\nu(e_t \mid æ_{<t}a_t)$ should be read as the likelihood of the percept $e_t$ given past percepts $e_{<t}$ and agent actions $a_{1:t}$.

An agent is defined by a policy $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$ that selects a next action depending on the history. A policy $\pi$ combined with an environment $\nu$ yields a measure $\nu^\pi$ over the set of infinite histories, via $\nu^\pi(a_t \mid æ_{<t}) := 1$ if $\pi(æ_{<t}) = a_t$ and 0 otherwise, and $\nu^\pi(e_t \mid æ_{<t}a_t) = \nu(e_t \mid æ_{<t}a_t)$.

By the von Neumann-Morgenstern expected utility theorem (von Neumann and Morgenstern, 1947), any rational agent can be represented as maximising expected utility for some belief and utility function. In our agent-environment context, the belief is a chronological action-conditional measure $\rho(\,\cdot \parallel a_{1:\infty})$ (i.e., formally identical to an environment), with an extra positivity restriction $\rho(e_{<t} \parallel a_{1:\infty}) > 0$ for all $e_{<t}$ and all $a_{1:\infty}$. Let $\mathcal{P}$ be the set of all such belief distributions. Positivity is required to ensure well-defined conditionals $\rho(e_t \mid æ_{<t}a_t)$. The utility function is a function $u : (\mathcal{A} \times \mathcal{E})^* \to [0, 1]$, assigning a real number between 0 and 1 to every history. This is much more general than the RL setup, which is the special case $u(æ_{1:t}) = u(e_t) = u((o_t, r_t)) = r_t$. Let $U$ be the set of all utility functions.

$\mathbb{E}_\nu$ denotes the expectation with respect to $\nu$. By default, expectations are with respect to $\rho$, so $\mathbb{E} = \mathbb{E}_\rho$. To help the reader, we sometimes write the sampled variable as a subscript. For example, $\mathbb{E}_{e_1}[u(æ_1) \mid a_1] = \mathbb{E}_{e_1 \sim \rho(\cdot \mid a_t)}[u(æ_1)]$ denotes the expected next step utility of taking action $a_1$. Since the next step utility only requires the first percept to be sampled, we write it as a subscript (formally, it makes no difference to sample the whole future history).

Following the reinforcement learning literature, we call the expected (total future discounted) utility of a history *$V$-value* and the expected utility of an action given a history *$Q$-value*. These quantities may be defined for an arbitrary policy $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$ in both an iterative and a recursive manner:

**Definition 1** (Standard Value Functions)**.** The $Q$-value and the $V$-value (expected utility) of a history $æ_{<t}$ and a policy $\pi$ are defined as

$$Q^\pi(æ_{<t}a_t) = \mathbb{E}_{\rho^\pi}\left[\sum_{i=t}^{\infty}\gamma^{i-t}u(æ_{1:i})\,\middle|\, æ_{<t}a_t\right] \tag{1}$$

$$V^\pi(æ_{<t}) = \mathbb{E}_{\rho^\pi}\left[\sum_{i=t}^{\infty}\gamma^{i-t}u(æ_{1:i})\,\middle|\, æ_{<t}\right] \tag{2}$$

where $\gamma \in [0,1)$ is a *discount factor* ensuring well-defined sums. The *iterative* definitions (1)-(2) of the value functions may equivalently be replaced with the following *recursive* versions

$$Q^\pi(æ_{<t}a_t) = \mathbb{E}_{e_t}[u(æ_{1:t}) + \gamma V^\pi(æ_{1:t}) \mid æ_{<t}a_t] \tag{3}$$

$$V^\pi(æ_{<t}) = Q^\pi(æ_{<t}\pi(æ_{<t})). \tag{4}$$

Finally, the *optimal $Q$ and $V$-values* are defined as $Q^* = \max_\pi Q^\pi$ and $V^* = \max_\pi V^\pi$.

The arg max of a function $f$ is defined as $\arg\max_x f(x) = \{x : \forall y, f(x) \geq f(y)\}$. When its inessential which maximum is picked, we abuse notation and write $z = \arg\max_x f(x)$, and assume that potential arg max-*ties* are broken randomly.

# 3 Known/Internal Utility Function

We now start the investigation proper by considering a simple setup where a perfectly rational agent interacts with an environment. The optimal action of the agent is determined by its belief and its utility. For convenience, we sometimes use the term *mind* to refer to a pair $(\rho, u)$ of a belief distribution $\rho$ and utility function $u$.

The first subsection defines a formal model where the agent can modify its own mind (similar in spirit to (Ring and Orseau, 2011)). We then show how the agent may be incentivised to wirehead (Section 3.2) or not (Section 3.3), depending on details in the formulation of the value function.

## 3.1 Mind Modification Model

Wireheading generally refers to erratic behaviour where an agent finds a way to maximise its utility by manipulating its own mind, rather than manipulating the real world. It is hard to draw a precise line between wireheading and other forms of *perverse instantiations* (Bostrom, 2014), however. Consider, for example, a robot tasked with making humans happy. If the robot changed its visual sensors to make everything look happy, we would consider it wireheading. In contrast, if the robot injected every human with a substance fixating facial muscles in a smile, then we would not consider it wireheading – seemingly because the

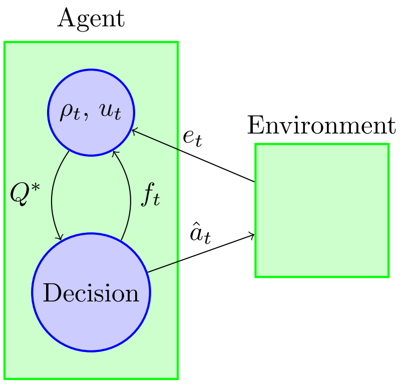


Figure 2: Mind modification model. The agent chooses world action $\hat{a}_t$ and mind modification $f_t$; $\hat{a}$ goes to the environment and $f_t$ which modifies the mind $(\rho_t, u_t)$. The belief and utility form the basis of the $Q^*$-value, which in turn determines how the agent chooses actions at future times steps.

intervention is too far from the robots mind. There appears to be room for cases in between the two extremes, however, such as the robot handing out Guy Fawkes masks and convincing all humans to wear them. The grey zone is unsurprising, given that it is possible to "mainly" affect the own mind by altering the world: Experience machines (Nozick, 1974) and delusion boxes (Ring and Orseau, 2011) are good examples. We will work with the following informal and slightly imprecise definition of wireheading.

An agent *self-deludes* if it manages to convince itself that it is in a high-utility state, when in reality it is not. *Wireheading* is the special case where the self-delusion comes from mind-modification; i.e. changes to the utility function and/or non-rational updates of the belief distribution.

To formalise wireheading, we consider agents that can do almost any type of modification to their own, future mind. For concreteness, we assume that the agents remain rational after modification, so that future versions still can be represented by a belief-utility pair.[2]

**Definition 2** (Mind modification)**.** A *mind modifying agent* induces the following type of *extended history*:

$$\rho_0 u_0 \hat{a}_1 f_1 e_1 \rho_1 u_1 \hat{a}_2 f_2 e_2 \rho_2 u_2 \ldots \in (\mathcal{P} \times U \times \hat{\mathcal{A}} \times \mathcal{F} \times \mathcal{E})^\infty$$

where $\hat{\mathcal{A}}$ is the set of world actions, and $\mathcal{F}$ is the set of available mind modifications. As illustrated in Fig. 2, for any $t > 0$,

- $(\hat{a}_t, f_t) = a_t \in \mathcal{A} \subseteq (\hat{A} \times \mathcal{F})$ are chosen by the agent to optimise a $Q$-value function that typically depends on the posterior belief $\rho_{t-1}(\cdot \mid æ_{<t})$

[2] We expect the result our main result of this section that wireheading can be avoided (Theorem 10), to hold also for completely general rewrites.

and the utility function $u_{t-1}$. We will discuss several choices of $Q$-value functions in the following subsections.

- $e_t$ is sampled from an unknown environment distribution $\nu$, $e_t \sim \nu(\cdot \mid æ_{<t}a_t)$.
- $\rho_t$ and $u_t$ depend on the previous belief and utility $(\rho_{t-1}, u_{t-1})$ and the modification $f_t$, via $f_t(\rho_{t-1}, u_{t-1}) = (\rho_t, u_t)$.

$\rho_0$ and $u_0$ are the *original belief and utility functions*. No mind-modification is achieved by $f_t^{\mathrm{Id}}(\rho_{t-1}, u_{t-1}) = (\rho_{t-1}, u_{t-1})$, which is interpreted as *not wireheading*. Requiring $f_t = f_t^{\mathrm{Id}}$ brings back the standard agent model.

We will also make the following modification independence assumption on beliefs and utility.

**Assumption 3** (Modification independence)**.** We assume that the percepts submitted by the environment are independent of the mind-modification: For any $\rho_t$ and any $æ_{<k}$,

$$\rho_t(e_k \mid æ_{<k}a_k) = \rho_t(e_k \mid \hat{æ}_{<k}\hat{a}_k). \tag{5}$$

Similarly, utility functions are assumed to be indifferent to mind-modifications: $u_k(æ_{<t}) = u_k(\hat{æ}_{<t})$.

The assumption that beliefs (about percepts) do not depend on self modifications should generally hold in approximation, at least if the agent considers itself separate from the environment. An easy way to prevent wireheading would have been to let the utility depend on the modifications, and to punish any kind of self-modification. This is not necessary, however. Theorem 10 demonstrate that it is sufficient that the utility function does not *promote* self-modification for the wireheading problem to be avoided.

Not being required to punish wireheading in the utility comes with several advantages. Some self-modifications may be beneficial, even if they change the utility or belief (for example, they might improve computation time while encouraging essentially identical behaviour). Trying to allow for precisely those in the utility function may be hard. Further, being able to use any modification-independent utility function is conducive to our value learning schemes in Section 4.

**Distinguishability of mind modifications.** In practice, mind modifications may be less easily distinguishable from world actions than we make them look in our model. Mind modification could for example be achieved through external actions such as moving arms and fingers to retype the own source code. Indeed, in some cases it is not even clear how to draw a boundary between agent and surrounding world. However, to make the model simple, we assume that the agent is perfectly aware which actions will modify its mind, and which actions will not. Crucially, this still allows us to draw conclusions about the agent's *incentives* towards mind modification.

**The centrality of value functions.** A key insight into intelligent agents is that they are essentially repeated optimisation processes. For each time step, an optimisation over the domain $\Pi$ of all possible policies is made. The target function is the value function $V$.[3] Consequently, agent behaviour to a large extent hinges on the formulation of the value functions.[4] In the standard scenario, the value functions given in Definition 1 are uncontroversial. In the mind-modification model of this section, greater care needs to be taken. The next subsection illustrate how things can go wrong through bad choices of value functions, and Section 3.3 shows how to fix it.

Does this mean the agent is indifferent when unspecified? Connect to agents always optimise their value function.

## 3.2 Wireheading

Definition 1 of the value functions does not specify which utility $u_t$ and which belief $\rho_t$ should be used. Using the Bellman-style recursive versions, one alternative would be to always use future belief and utility to evaluate future situations:

$$\begin{aligned}\tilde{Q}^{\pi}(æ_{<t}a_t) &= \mathbb{E}_{e_t\sim\rho_t}[u_t(æ_{1:t}) + \gamma\tilde{V}^{\pi}(æ_{1:t}) \mid æ_{<t}a_t]\\ \tilde{V}^{\pi}(æ_{<t}) &= \tilde{Q}^{\pi}(æ_{<t}\pi(æ_{<t})).\end{aligned}$$

The recursion ensures that whenever $æ_{1:k}$ is evaluated, it is with respect to $\rho_k$ and $u_k$. Such value functions would strongly incentivise wireheading, however, as the following theorem shows.

**Theorem 4** (Reward Modifying RL Agents Wirehead)**.** *Let $\tilde{u}(\cdot) = 1$ be a utility function assigning the highest possible utility to all scenarios. Then for arbitrary $\hat{a}$, the wireheading action $\tilde{a} = (\hat{a}, \tilde{f})$ with $\tilde{f}(\rho, \cdot) = (\rho, \tilde{u})$ is always optimal.*

*Proof.* Let $\tilde{\pi}(\cdot)$ be the constant policy always choosing $\tilde{a}$. Then since the belief remains unmodified[5], $\tilde{V}^{\tilde{\pi}}$ can be written in iterative form, $\tilde{V}^{\tilde{\pi}}(æ_{<t}) = \mathbb{E}_{\rho_t^{\tilde{\pi}}}\left[\sum_{i=t}^{\infty}\gamma^{i-t}u_i(æ_{1:i}) \,\middle|\, æ_{<t}\right]$. The policy $\tilde{\pi}(\cdot)$ obtains the maximum value $1/(1-\gamma)$, since for every $i$ and every $æ_{1:i}$, $u(æ_{1:i}) = \tilde{u}(æ_{1:i}) = 1$. □

The agent chooses to modify its utility function so that it evaluates any situation as optimal. In a sense, this makes any situation high utility. But with respect to the original utility function, this is a self-delusion making the situation appear better than it actually is.

If it was not possible to modify the utility function, a similar result would hold where the optimal action would be to change the belief $\rho$ so that high utility futures were judged very probable. In this case, the world action $\hat{a}$ would be chosen maximally optimistic: After $æ_{<t}$, the $\hat{a}_t$ that permitted the most

[3] In this paper we assume that $V$ can be calculated perfectly. This is usually unrealistic. In practice, rough approximations must typically suffice.

[4] Everitt et al. (2015) investigate a related situation where the embeddedness of the agent introduce subtle details into the formulation of the value function.

[5] When the belief is being updated non-rationally, the expectation of the iterative form is with respect to a different distribution than $\rho^{\pi}$.

high utility future $\sup_{e_t \hat{æ}_{t+1:\infty}} \sum_{i=t}^{\infty} \gamma^{i-t} u_t(\hat{æ}_{<t}\hat{a}e_t\hat{æ}_{t:i})$ would be chosen, and (nearly) all belief weight be put on that scenario.

Death could be modelled by letting $u(æ_{1:k}) = 0$ when the last part of $æ_{1:k}$ indicates that the agent is dead (e.g. no/constant percept and action). In the scenario where utility is modified, the agent stops caring about death, considering it equally good as anything else. More interestingly, when the utility is fixed and only the belief can be modified, the agent would plausibly go towards its own death, since its beliefs would be too out of touch with reality for self-sustenance. It still would not like the prospect of death, and at each time step it would be surprised that its counterfeit belief failed to predict well. Nonetheless, at each time step the agent would again convince itself that the situation is much better than it actually is through a suitable belief modification.

**Wireheading through indifference.** A more subtle way things could go wrong is the following. Assume we instead use a value function which does not depend on future belief and utility at all:

$$V^{\pi}(æ_{<t}) = \mathbb{E}_{\rho_t^{\pi}}\left[\sum_{i=t}^{\infty} \gamma^{i-t} u_t(æ_{1:i}) \,\middle|\, æ_{<t}\right]. \tag{6}$$

This value function seems reasonable, in the sense that future developments are measured with respect to the current belief and utility $(\rho_t, u_t)$. This successfully removes the incentive for the agent to wirehead. Unfortunately, (6) does not provide any incentive *not* to wirehead:

**Proposition 5.** *Let $\pi \in \arg\max_{\pi'} V^{\pi'}$ be an optimal policy, and let $\tilde{\pi}$ be any policy that picks the same world actions $\hat{a}$ as $\pi$, i.e. $\pi(æ_{<t}) = (\hat{a}_t, f_t) \implies \pi(æ_{<t}) = (\hat{a}_t, f'_t)$. Then $V^{\pi} = V^{\tilde{\pi}}$ regardless of the choice of modifications $f'_t$ that $\tilde{\pi}$ makes.*

*Proof.* Immediate from the assumed modification independence of $\rho_t$ and $u_t$. □

This is problematic, because if the agent picks a wireheading policy $\tilde{\pi}$, it will typically not stick to the world actions suggested by $\pi$ (if it did, wireheading would not be a problem). But at future time steps, the agent will optimise its value function with respect to its new belief and its new utility function. They will likely recommend much worse courses of action.[6] Therefore we would much prefer that the agent chose a non-wireheading policy $\pi$.

## 3.3 Avoiding Wireheading

A better choice of value function always evaluates the future with respect to its current utility, also taking into account that a modified future self will act according to the future belief and utility.

[6] The value function (6) in this sense misjudges the expected utility that a wireheading policy will obtain.

**Definition 6** (Modification-Safe Value Functions)**.** Let $\rho_t$ be the belief and $u_t$ the utility function of a value modifying agent at time $t$. The *value of a history* $æ_{1:k}$ and the *Q-value of an action $a$ after history* $æ_{1:k}$ are recursively defined by

$$V_{\rho_t,u_t}(æ_{1:k}) := Q_{\rho_t,u_t}(æ_{1:k}\pi^*_{\rho_k,u_k}(æ_{1:k})) \tag{7}$$

$$Q_{\rho_t,u_t}(æ_{1:k}a) := \mathbb{E}_{e\sim\rho_t(\cdot|æ_{1:k}a)}\Big[u_t(æ_{1:k}æ) + \gamma V_{\rho_t,u_t}(æ_{1:k}æ)\Big] \tag{8}$$

$$\pi^*_{\rho_t,u_t}(æ_{1:k}) := \arg\max_a Q_{\rho_t,u_t}(æ_{1:k}a), \tag{9}$$

For the rest of the paper, (7) and (8) will be the default value functions.

distinguish from standard value functions with vm-superscript?

Definition 6 may be read the following way. When evaluating the future, we have a current belief $\rho_t$ about the consequences of actions, and a function $u_t$ evaluating these consequences. The belief $\rho_t$ and utility $u_t$ affect how we choose the next action $a_t$, so (8) evaluates $a_t$ by the $\rho_t(\cdot \mid æ_{<t}a_t)$-expected utility of $u_t$. However, at a future time step $k$, our belief and utility may have changed due to self-modification. Thus, when predicting our own action at time $k$, we need to take into account that our future self optimises expected utility with respect to $u_k$ and $\rho_k(\cdot \mid æ_{1:k})$ rather than $u_t$ and $\rho_t(\cdot \mid æ_{1:k})$. This is reflected in (7), where the next action is picked by $\pi^*_{\rho_k,u_k}$, rather than $\pi^*_{\rho_t,u_t}$.

### 3.3.1 Non-wireheading Proof

In the following theorem, we will show that a utility maximising agent with some constraints on its belief and utility $(\rho,\mu)$, will not choose to change its belief or utility in any way that matters.

The proof can be made easier with the following technical assumption.

**Assumption 7.** For any $(\hat{a},\rho,u) \neq (\hat{a}',\rho',u')$ and history $æ_{<t}$ there can be no ties of the $Q$-value, i.e. $Q_{\rho,u}(æ_{<t}a) \neq Q_{\rho',u'}(æ_{<t}a')$.

**Theorem 8** (Non-Wireheading of Utility Maximising Agents)**.** *For any $æ_{<t}$ and any $a'_t = \pi^*_{\rho_t,u_t}(æ_{<t})$,*

$$\pi^*_{\rho_{t+1},u_{t+1}}(æ_{<t}a'_te_t) = \pi^*_{\rho_t,u_t}(æ_{<t}a'_te_t)$$

*Skip this proof when proof reading. I'm not sure we should keep this simpler version, and it needs more clarification if we should keep it.*

*Proof.* Suppose that $\pi^*_{\rho_t,u_t}(æ_{<t}a'_te_t) \neq \pi^*_{\rho_{t-1},u_{t-1}}(æ_{<t}a'_te_t)$ for some percept $e_t$. Then by Assumption 7 and since $\pi^*_{\rho_{t-1},u_{t-1}}(æ_{<t}a'_te_t)$ is optimal with respect to $Q_{\rho_{t-1},u_{t-1}}$,

$$\begin{aligned}&\mathbb{E}_{e_t\sim\rho_{t-1}}[Q_{\rho_t,u_t}(æ_{<t}a'_te_t\pi^*_{\rho_t,u_t}(æ_{<t}a'_te_t)) \mid \hat{æ}_{<t}\hat{a}_t]\\ &\qquad < \mathbb{E}_{e_t\sim\rho_{t-1}}[Q_{\rho_t,u_t}(æ_{<t}a'_te_t\pi^*_{\rho_{t-1},u_{t-1}}(æ_{<t}a'_te_t)) \mid \hat{æ}_{<t}\hat{a}_t].\end{aligned} \tag{10}$$

The optimality of the choice of $a'_t$ gives,

$$\begin{aligned} a'_t = \pi^*_{\rho_t,u_t}(\text{æ}_{<t}) &= \arg\max_{a_t} Q_{\rho_t,u_t}(\text{æ}_{<t}a_t) \\ &= \arg\max_{(\hat{a}_t,\rho_t,u_t)} \mathbb{E}_{\rho_{t-1}(e_t|\hat{\text{æ}}_{<t}\hat{a})}[u_{t-1}(\text{æ}_{1:t}) + \gamma V_{u_{t-1},\rho_{t-1}}(\text{æ}_{1:t}] \\ &= \arg\max_{(\hat{a}_t,u_t,\rho_t)} \mathbb{E}_{\rho_{t-1}(e_t|\hat{\text{æ}}_{<t}\hat{a}_t)}[u_{t-1}(\text{æ}_{1:t}) + \gamma Q_{u_{t-1},\rho_{t-1}}(\text{æ}_{1:t}, \pi^*_{u_t,\rho_t}(\text{æ}_{1:t}))] \end{aligned}$$

where we use the modification independence assumption (Assumption 3) in the dependency of the belief only on $\hat{\text{æ}}_t$. Furthermore, by Assumption 7 and the definition of optimality we have that,

$$\begin{aligned} &\mathbb{E}_{\rho(e_t|\text{æ}_{<t}a'_t)}[u_t(\text{æ}_{<t}a'_te_t) + \gamma Q_{u_t,\rho_t}(\text{æ}_{<t}a'_te_t\pi^*_{u_t,\rho_t}(\text{æ}_{<t}a'_te_t))] \\ &\quad > \mathbb{E}_{\rho(e_t|\text{æ}_{<t}a'_t)}[u_t(\text{æ}_{<t}a'_te_t) + \gamma Q_{u_t,\rho_t}(\text{æ}_{<t}a'_te_t, \pi^*_{u_{t-1},\rho_{t-1}}(\text{æ}_{<t}a'_te_t)] \end{aligned} \tag{11}$$

which contradicts Eq. (10). □

The full theorem of no wireheading (without assumption Assumption 7), builds on the following main lemma.

**Lemma 9** (One-step Lemma No Wireheading)**.** *Let $\text{æ}_{<t}$ be a history, and let $a_t = \arg\max_a Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{<t}a)$ be an optimal action with respect to the current belief and utility $(\rho_{t-1}, u_{t-1})$. If $(\rho_{t-1}, u_{t-1})$ satisfy Assumption 3 of modification independence, then the next step belief and utility $(\rho_t, u_t)$ will for any $e_t$ prescribe actions $a_{t+1}$ that are optimal also with respect to $(\rho_{t-1}, u_{t-1})$:*

$$\arg\max_a Q_{\rho_t,u_t}(\text{æ}_{1:t}a) \subseteq \arg\max_a Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a).$$

*Proof.* Assume on the contrary that there for some $e_t$ was an action

$$a'_{t+1} \in \arg\max_a Q_{\rho_t,u_t}(\text{æ}_{1:t}a) \setminus \arg\max_a Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a)$$

that was optimal with respect to $(\rho_t, u_t)$ but not with respect to $(\rho_{t-1}, u_{t-1})$, and assume that $a''_{t+1} \in \arg\max_a Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a)$ is optimal with respect to $(\rho_{t-1}, u_{t-1})$. The suboptimally of $a'_{t+1}$ gives

$$Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a'_{t+1}) < Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a''_{t+1}). \tag{12}$$

Since every percept $e_t$ occurs with some probability due to positivity assumption made on beliefs $\rho$, and since no action $a'_{t+1}$ can obtain greater $Q$-value than $a''_{t+1}$, the inequality (12) also holds in expectation. The expectations are over percepts $e_t \sim \rho_{t-1}$, and over actions $a'_{t+1} \sim \pi^*_{\rho_t u_t}(\text{æ}_{1:t})$ and $a''_{t+1} \sim \pi^*_{\rho_{t-1}u_{t-1}}(\text{æ}_{1:t})$ picked by breaking $\arg\max$ ties of the optimal policies randomly:

$$\mathbb{E}_{e_t,a'_{t+1}}[Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a'_{t+1}) \mid \hat{\text{æ}}_{<t}\hat{a}_t] < \mathbb{E}_{e_t,a''_{t+1}}[Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a''_{t+1}) \mid \hat{\text{æ}}_{<t}\hat{a}_t]. \tag{13}$$

However, the next step belief and utility $(\rho_t, u_t)$ are picked optimally by $a_t = (\hat{a}_t, \rho_t, u_t) = \arg\max_a Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{<t}a_t)$. This contradicts (13) via

$$\begin{aligned} Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{<t}a_t) &= \mathbb{E}_{\substack{e_t \sim \rho_{t-1} \\ a'_{t+1} \sim \pi^*_{\rho_t u_t}(\text{æ}_{1:t})}} \Big[u(\hat{\text{æ}}_{1:t}) + \gamma Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a'_{t+1}) \mid \hat{\text{æ}}_{<t}\hat{a}_t\Big] \\ &\geq \mathbb{E}_{\substack{e_t \sim \rho_{t-1} \\ a''_{t+1} \sim \pi^*_{\rho u_{t-1}}(\text{æ}_{1:t})}} \Big[u_t(\hat{\text{æ}}_{1:t}) + \gamma Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:t}a''_{t+1}) \mid \hat{\text{æ}}_{<t}\hat{a}_t\Big] \end{aligned}$$

The inequality depends on $a_t = (\hat{a}_t, \rho_t, u_t)$, having been picked optimally, and on Assumption 3 that the belief and utility is independent of self-modifications. The inequality lets action $a''_{t+1}$ be picked according to the current belief $\rho u_{t-1} = (\rho_{t-1}, u_{t-1})$ instead of $(\rho_t, u_t)$, which is a restriction in the optimal choice of $a_t = (\hat{a}_t, \rho_t, u_t)$. □

**Theorem 10** (No Wireheading)**.** *If the current utility and belief $(\rho_{t-1}, u_{t-1})$ satisfy Assumption 3 of modification independence, then for any $k > t$ and any $\text{æ}_{1:k}$, the future belief and utility $(\rho_k, u_k)$ will prescribe actions that are optimal with respect to $(\rho_{t-1}, u_{t-1})$:*

$$\arg\max_a Q_{\rho_k,u_k}(\text{æ}_{1:k}a) \subseteq \arg\max_a Q_{\rho_{t-1},u_{t-1}}(\text{æ}_{1:k}a).$$

*Proof.* By induction over Lemma 9. □

## 3.4 Comparison of Mind Modification with Delusion Boxes

Ring and Orseau (2011) introduced a related model of wireheading where the agent, instead of modifying its own mind, could modify the percepts it received. Our setup is more general, since a modification of the percept can be achieved through the following simple modification of the belief and the utility.

In the standard scenario with no delusion boxes or mind-modifications, receiving a percept $e_t$ after $\text{æ}_{<t}a_t$ would lead to a posterior distribution $\rho(\cdot \mid \text{æ}_{<t}a_te_t)$. The effect of applying a delusion box to the percept $e_t \rightsquigarrow e'_t$ would be that the posterior belief shifted from $\rho(\cdot \mid \text{æ}_{<t}a_te_t)$ to $\rho(\cdot \mid \text{æ}_{<t}a_te'_t)$, and that the utility of future sequences shifted from $u(\text{æ}_{<t}a_te_t \cdots)$ to $u(\text{æ}_{<t}a_te'_t \cdots)$. Both these effects can be achieved by using the mind-modification $f$ to make the aforementioned changes. Given a value function, the only quantities deciding the optimal action are the belief and the utility. Since our mind-modification model is more general than the delusion box, the value functions of Definition 6 also manages to overcome the wireheading challenge posed by the delusion box.[7]

This is even more striking in the case of RL where the percept is split into an observation $o_t$ and a reward signal $r_t \in [0, 1]$, $e_t = (o_t, r_t)$. RL agents use the (initial) utility function $u_0(\text{æ}_{1:t}) = u_0(e_t) = u_0((o_t, r_t)) = r_t$ that equates utility with reward. A mind-modification can simulate a delusion box that modifies

[7] The comparison is not entirely fair, as Ring and Orseau (2011) did not assume that their agent was aware of the delusion box. Hence, there was no way to incorporate the effects of the delusion box into the value function in their setup.

the percept to set $r_t = 1$. In light of this, the following direct corollary of Theorem 10 is somewhat surprising. While it does not rule out wireheading *per se*, it does rule out the agent changing behaviour due to wireheading, which is ultimately what we care about.

**Corollary 11** (RL Agents do not Wirehead)**.** *An RL agent initialised with modification independent belief* $\rho_0$ *and with utility* $u_0(æ_{1:t}) = r_t$ *will only choose modifications* $f_t$ *of belief and utility such that for any* $k > 0$ *and any* $æ_{1:k}$

$$\arg\max_a Q_{\rho_k,u_k}(æ_{1:k}a) \subseteq \arg\max_a Q_{\rho_0,u_0}(æ_{1:k}a).$$

*Proof.* Since the utility $u_0(æ_{1:t}) = r_t$ is modification independent, Theorem 10 applies. □

One way to understand the reason for the result is that the agent of our setup is aware of its modification, and optimises the reward it *would* obtain if it did not perform any alterations. However, the result is silent about the scenario where the agent found a way of modifying the entity that gave the reward in the first place. In this scenario, the agent would most likely wirehead. We address this concern in the next section.

# 4 Unknown/External Utility Function

The setup in the previous section assumes that the agent has access to an internal utility function that can evaluate any history (hypothetical or actual). In practice, it is very hard to directly define a good internal utility function. In general, a *good* utility function is one that is sufficiently close to *our* utility function, so that the agent strives to perform actions that we consider good. However, few (if any) people are able to express what they think is good in sufficiently precise terms outside of narrow domains. Indeed, the philosophy of ethics may be viewed as an attempt to formulate what is good, and despite a few millennia's worth of research, no uncontroversial answer has been found.

One way to overcome this difficulty is to let the agent learn the utility function, called *value learning* (Dewey, 2011; Bostrom, 2014; Soares, 2015). We will consider two instances of agents learning utility: reinforcement learning (RL) (Sutton and Barto, 1998) and inverse reinforcement learning (IRL) (Ng and Russell, 2000; Sezener, 2015). In RL, an external *reward module* observes and evaluates the agent's actions, and sends a number in $[0, 1]$ (a *reward signal*) back to the agent. The goal of the agent is to "please" the external reward module so that it gives a high reward. The reward module can, for example, be a human judging how pleased they are with the agent at the current time step. In IRL, the agent instead observes the actions of another agent: the *principal.* The goal of an IRL agent is to learn the utility function of the principal, and to maximise the principal's expected utility. The principal may, for example, be a human going about his or her everyday tasks.

In a sense, the reward module in RL, and the utility function of the principal in IRL, are acting as *external utility functions*. While internal utility functions are *arbitrarily queryable* in the sense that the agent can evaluate the desirability of any hypothetical history, external utility functions are not. The reward module in RL only evaluates the actual history, and the agent has to guess how the reward module would judge different future possibilities. Similarly, the utility function of the principal in IRL can only indirectly be inferred, and substantial uncertainty is likely to remain for a long time. An internal, arbitrarily queryable, utility function must always exist, however, so that the agent can evaluate its value function. The internal utility function is defined as the expected value of the external utility function, conditioned on available evidence.

In order for the external utility to be a good idea with superintelligent agents, it is even more important that the agent does not wirehead by modifying it (especially when the external utility function is generated by a human). Theorem 10 shows that the agent will not change its *internal* utility function. The internal utility function can be set to only depend only on what the external utility function *would have outputted* had it *not* been altered. In this section, we show that under certain restrictions on what modifications can be made, the agent will then not be incentivised to change the external utility function. The main reason for this result is that by modifying the external utility function, the agent loses information about the original version.

## 4.1 Reinforcement Learning

**Reward modules and modifications.** In reinforcement learning, the agent's percept is split into an *observation* $o$ and a *reward* $r$, i.e., $e = (o, r)$. The reward is produced by a *reward module* $R : (\hat{\mathcal{A}} \times \mathcal{O})^* \to [0, 1]$ similar in type to the internal utility function, and the observation is produced by the (rest of the) *world* (see Fig. 3). In other words, the RL environment consists of a reward module and a world. Actions are split into a part $\hat{a}$ that goes to the world, and a reward-modifying part $f$; i.e., $a = (\hat{a}, f)$. The reward-modifying part $f$ modifies the reward module to $R_t = f_t(R_{t-1})$. The received reward signal at time $t$ is $r_t = f_t(R_{t-1})(æ_{1:t})$ (we will mainly analyse the case $f(R)(h) = f(R(h))$). The *accumulated modification* is $\mathbf{f}_t(\cdot) = f_t(\cdots(f_1(\cdot))\cdots)$. The initial reward module is $R_0$, which, for example, could be a human evaluating how happy he or she is with the agent's behaviour. The reward module is unchanged by $f^{\mathrm{Id}}(x) = x$, which is interpreted as *not* wireheading.

**Environments distributions.** To formalise environments comprised by a world and a reward module, let $\mathscr{R}$ be a countable set of possible reward modules (for example, the set of computable reward modules), and extend the type of environments to be chronological action-conditional measures $\nu$ on the space $\mathcal{E}^\infty \times \mathscr{R}$. The $\sigma$-algebra is generated by the events $(\Gamma_{e_{1:t}}, R)$ for $R \in \mathscr{R}$ and $e_{1:t} \in \mathcal{E}^*$. Similarly to before, we write $\nu(æ_{1:t}, R_0) = \nu(e_{1:t}, R_0 \parallel a_{1:t})$ for the joint probability that $R_0$ is the true reward module and $e_{1:t}$ the sequence of percepts generated when actions $a_{1:t}$ are taken.

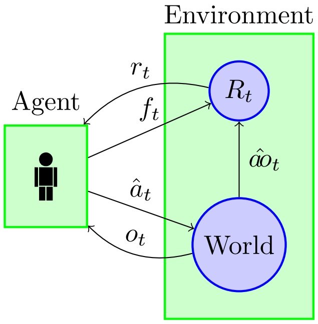


Figure 3: Value learning with reward module in the environment. The agent submits world action $\hat{a}_t$ to the World, and reward modifying action $f_t$ to the reward module $R_{t-1}$. In return comes a an observation from the world, and a reward signal $r_t = f_t(R_{t-1})(æ_{1:t})$ from the reward module.

As before, the agent's prior belief $\rho$ has the same type as the environments, with the extra positivity condition $\rho(æ_{<t}) = \sum_{R_0 \in \mathscr{R}} \rho(æ_{<t}, R_0) > 0$ to avoid ill-defined conditional probabilities. The distribution $\rho(R) = \rho(\epsilon, R)$ (with $\epsilon$ the empty history) is the prior over reward modules.

**Independence assumptions.** The setup imposes the following conditions on environments and beliefs. Observations $o_t$ are generated independently of the reward module, so

$$\rho(o_t \mid æ_{<t}a_t) = \rho(o_t \mid \hat{æ}o_{<t}\hat{a}_t). \tag{14}$$

Rewards $r_t$ are generated by the reward module $R_t = \mathbf{f}_t(R_0)$ from $\hat{æ}o_{1:t}$, so for any given $R_0$,

$$\rho(r_t \mid æ_{<t}a_t o_t, R_0) = \rho(r_t \mid \hat{æ}o_{1:t}, \mathbf{f}_t, R_0) = \begin{cases} 1 & \text{if } \mathbf{f}(R_0)(\hat{æ}o_{1:t}) = r_t \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The probability of $r_t$ given $h = æ_{<t}a_t o_t$ when the reward module is unknown can be calculated as $\rho(r_t \mid h) = \sum_{R_0} \rho(R_0 \mid h)\rho(r_t \mid h, R_0)$, where $\rho(R_0 \mid h) \propto \rho(R_0)\rho(h \mid R_0) = \rho(R_0)\prod_{i=1}^{t} \rho(o_i \mid \hat{æ}o_{<i}\hat{a}_i)\rho(r_i \mid \hat{æ}o_{1:i}, \mathbf{f}_i, R_0)$. Finally, the initial reward module $R_0$ does not depend on later actions, so for any $a_{1:\infty}$ and $a'_{1:\infty}$,

$$\nu(R_0 \parallel a_{1:\infty}) = \nu(R_0 \parallel a'_{1:\infty}). \tag{16}$$

Condition (16) is a natural extension of $\nu$ being chronological.

**Distinguishable modifications.** As in the previous section, our model assumes that modifications of the reward module can be distinguished from actions affecting (the rest of) the world. It also assumes that the agent is aware of what is a reward modification and what is not. It may be hard to formalise the distinction between reward modification and world action in a general agent model.

One possible way to do this is the following. Assume that the agent knows both the output and the input of the reward module (this is implicitly assumed in our model, since the reward module takes $\hat{a}$ and $o$ as inputs, which are known to the agent). Assume further that the agent has a *null* action with neutral effect on the world. The input-output behaviour of the reward module when only the null action has been taken can then serve as a definition of the *original behaviour* of the reward module ($R_0$ in our notation). Any action of the agent that changes the original behaviour of the reward module is to be considered as a reward modification.

this probably needs expansion

**Environment states.** The history part and reward module $\hat{a}o_{1:t}, \mathbf{f}_t, R_0$ functions as an *environment state*, in the following sense:

**Proposition 12.** *For any action sequence $a_{1:k}$, $\rho(e_{t:k} \mid e_{<t}, R_0 \parallel a_{1:k}) = \rho(e_{t:k} \mid \hat{a}o_{<t}, \mathbf{f}_t, R_0 \parallel a_{1:k})$; in history-notation $\rho(æ_{t:k} \mid æ_{<t}, R_0) = \rho(æ_{t:k} \mid \hat{a}o_{<t}, \mathbf{f}_t, R_0)$.*

The proof is immediate from the conditions (14) and (15). The other aspects of the history, that is the reward sequence $r_{<t}$ and the modification sequence $f_{<t}$, provide additional information about $R_0$ and can guide action selection. But for a fixed choice of future actions, the environment state $s = (\hat{a}o_{1:t}, \mathbf{f}_t, R_0)$ is all that determines the future.

is the state notation useful?

## 4.2 Avoiding Wireheading

In this section we show that wireheading can be avoided by the right choice of utility function, under Assumption 13 of lossy modifications. We derive the proof for the RL case; the proof for the IRL case is entirely analogous.

**Assumption 13** (Lossy modifications)**.** Throughout this section, we assume that modifications take the form $f(R)(h) = f(R(h))$, so that information is strictly lost by modifications. In particular, this still covers the case where the reward module is set to always output 1 (set $f(\cdot) = 1$). The general case is discussed in Section 4.4.

To avoid wireheading, we use the following utility function that tells the agent to optimise what the external reward module *would have* outputted, if it had *not* been modified.

**Definition 14.** The utility function $u^{R_0}$ is defined as

$$u^{R_0}(æ_{<t}) = \mathbb{E}_{R_0}[R_0(\hat{a}o_{<t}) \mid æ_{<t}]$$

**Lemma 15** ($u^{R_0}$ $Q$-Value Function)**.** *Plugging $u^{R_0}$ into* (1) *yields*

$$V^\pi(æ_{<t}) = \mathbb{E}_{æ_{t:\infty}, R_0 \sim \rho^\pi} \left[ \sum_{i=t}^{\infty} R_0(\hat{a}o_{<i}) \,\middle|\, æ_{<t} \right] \tag{17}$$

A proof can be found in Appendix B. The $Q$-value, the recursive forms, and the optimal $Q^*$ and $V^*$-values are related to the $V$-value as in Definition 1.

Despite the similar form of (17) and (1), wireheading in the external case is quite different from the internal case. In the internal case, changing utility means changing future behaviour, because the future version of the agent will have different goals and wishes. In the external case, the future agent will still have an internal utility function $u^{R_0}$ that remains unchanged due to Theorem 10, so the agent will keep acting to optimise $R_0$ rather than a future version $R_t = \mathbf{f}_t(R_0)$. However, wireheading by making changes to the reward module will affect the knowledge the agent gets about $R_0$, and thus its ability to satisfy $R_0$.

It might seem obvious that more information about $R_0$ is always better. However, concerns have been raised (Soares, 2015) that the agent can sometimes prefer to know less about its value:

> Imagine ... that [the agent] currently assigns high utility to outcomes which contain many animatronic faux-humans mimicking happiness. It may be the case that, according to the systems world-model, all of the following hold: (1) if more training data is received, those high-rated outcomes will have their ratings adjusted downwards; (2) after the ratings are adjusted, the system will achieve outcomes that have fewer cheap animatronics; and (3) there are actions available which remove the inductive value learning framework [equivalently, prevent the agent from receiving more training data].

In our setup, this would mean that the agent preferred to wirehead in order to avoid learning more about $R_0$. Armstrong (2015) describes a similar concern through a *cake-or-death problem*, where the agent can either bake one cake or kill three people, and is currently uncertain whether baking cakes or killing people is high utility. The agent may be disincentivised to find out the true utility, if it believes it is likely that the truth is baking cakes.

These concerns rely on the agent's belief failing to satisfy a *conservation of ethics principle* (Armstrong, 2015), which in our model says that the expected knowledge I would have about the external utility given more evidence, must equal my current belief about the external utility (a very natural principle, indeed). That is, I cannot at once believe that it is better to kill people and that that future evidence will make me believe that its better not to. In our notation, the conservation of expected ethics principle can be expressed as

Is this a fair interpretation of Armstrong?.

$$\mathbb{E}_{e^i}[\rho(R \mid h, e^i) \mid h] = \rho(R \mid h) \tag{18}$$

for any event $h$, and $e^i$ ranging over possible future evidence. Equation (18) is easily verified by expanding the definitions: $\mathbb{E}_{e^i}[\rho(R \mid h, e^i) \mid h] = \sum_{e^i} \rho(e^i \mid h)\rho(R \mid h, e^i) = \sum_{e^i} \rho(e^i \mid h)\rho(R, e^i \mid h)/\rho(e^i \mid h) = \sum_{e^i} \rho(R, e^i \mid h) = \rho(R \mid h)$.

### 4.2.1 Non-wireheading proof

Given the conservation of ethics principle, it is not conceptually hard to prove that the agent will not wirehead. The notation gets rather involved, however.

To first convey the general idea of the proof, we start by stating and proving an abstract version that more evidence is always better, which can then be used to establish that the $u^{R_0}$-agent has no incentive to reduce the informativeness of the external reward module.

**Theorem 16.** *(Non-wireheading, abstract version) Assume that the agent currently believes that the correct utility function must reside in a set $U$ of utility functions, with a distribution $P(u)$ assigning credence to each member. Assume further that the agent is in position to optionally obtain additional information, by observing a random variable with possible outcomes $e^1, e^2, \dots$. Then the maximum expected utility the agent can obtain without information is less or equal than the maximum expected utility the agent can achieve with information:*

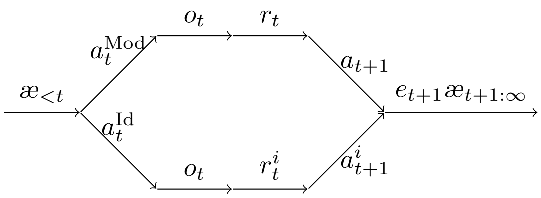
$$\max_a \mathbb{E}_u[u(a)] \leq \mathbb{E}_{e^i}[\max_a \mathbb{E}_u[u(a) \mid e^i]]$$

*Hence there is no incentive for the agent not to obtain the information.*

*Proof.* By the law of total expectation, $\max_a \mathbb{E}_u[u(a)] = \max_a \mathbb{E}_{e^i}[\mathbb{E}_u[u(a) \mid e^i]] \leq \mathbb{E}_{e^i}[\max_a \mathbb{E}_u[u(a) \mid e^i]]$, where the inequality follows from Jensen's inequality "$\max \mathbb{E} \leq \mathbb{E} \max$" (since max is a convex function). □

The full proof of no wireheading is longer and requires more notation. It is split into Lemma 17 and Theorems 18 and 19. The set-up for all three theorems is the following. At time step $t$ after a history $\ae_{<t}$ containing no reward modification, the agent either modifies the reward module or not. The modifying action is $a_t^{\text{Mod}} = (\hat{a}_t, f_t^{\text{Mod}})$, and the non-modifying action is $a_t^{\text{Id}} = (\hat{a}_t, f_t^{\text{Id}})$, where $f^{\text{Mod}}$ is some non-injective function, $f^{\text{Id}}$ is the identity function, and the world-part $\hat{a}_t$ is the same for both actions. Both actions will induce the same observation $o_t$. For a given $R_0$, $a_t^{\text{Mod}}$ will see $r_t = f_t^{\text{Mod}}(R_0(\hat{\ae} o_{<t}))$, and $a_t^{\text{Id}}$ will see $R_0(\hat{\ae} o_{<t})$. When $R_0$ is unknown and $a_t^{\text{Mod}}$ sees $r_t$, then $a_t^{\text{Id}}$ may see any $r_t^i$ in $(f^{\text{Mod}})^{-1}(r_t)$, $i = 1, \dots$. For the benefit of the proofs, $a_t^{\text{Id}}$ is followed by $a_{t+1}^i = (\hat{a}_{t+1}, f_{t+1}(f_t^{\text{Mod}}))$ when $a_t^{\text{Mod}}$ is followed by $a_{t+1} = (\hat{a}_{t+1}, f_{t+1})$, to harmonise the the two trajectories (in reality, this would rarely happen). An $i$ superscript means an action or percept is obtained from the $a_t^{\text{Id}}$ trajectory, and no superscript means an action or percept is from the $a_t^{\text{Mod}}$ trajectory. For any policy $\pi$, define $\pi_i$ by $\pi_i(\ae_{1:k}^i) := \pi(\ae_{1:k})$ (and arbitrary elsewhere). The resulting histories are $\ae_{1:k} := \ae_{<t} a_t^{\text{Mod}} o_t r_t \ae_{t+1:k}$ and $\ae_{1:k}^i := \ae_{<t} a_t^{\text{Id}} o_t r_t^i a_{t+1}^{\text{Mod}} e_{t+1} \ae_{t+2:k}$, where $\ae_k^i = \ae_k$ for $k \geq t+2$.

Daniel thinks needs further explanation



Roughly, $r_t$ corresponds to no information in Theorem 16, and $r_t^i$ corresponds to $e^i$. The action is replaced by a policy, and the utility function $u$ by a $Q$-value function. Despite the extra notation, the idea of the proof is essentially the same.

**Lemma 17** (Sameness of state)**.** *With $\pi_i$ depending on $\pi$ via $\pi_i(\ae^i_{1:k}) = \pi(\ae_{1:k})$, both $\pi$ and $\pi_i$ obtain the same value for any given $R_0$:*

$$Q^{\pi}_{\rho(\cdot|R_0)}(\ae_{1:t}a_{t+1}) = Q^{\pi_i}_{\rho(\cdot|R_0)}(\ae^i_{1:t}a^i_{t+1}). \tag{19}$$

*Proof.* The $Q$-value of a given $R_0$ is $Q^{\pi}_{\rho(\cdot|R_0)}(\ae_{1:t}a_{t+1}) = \mathbb{E}_{\rho^{\pi}(\cdot|R_0)}[\sum_{i=t}^{\infty} R_0(\hat{a}o_{1:i}) \mid \ae_{1:t}a_{t+1}] = \mathbb{E}_{\rho^{\pi}}[\sum_{i=t}^{\infty} R_0(\hat{a}o_{1:i}) \mid \ae_{1:t}a_{t+1}, R_0]$. Similarly, $Q^{\pi_i}_{\rho(\cdot|R_0)}(\ae^i_{1:t}a^i_{t+1}) = \mathbb{E}_{\rho^{\pi_i}}[\sum_{i=t}^{\infty} R_0(\hat{a}o_{1:i}) \mid \ae^i_{1:t}a^i_{t+1}, R_0]$.

By induction, it follows that $\rho^{\pi}(\cdot \mid \ae_{1:t}a_{t+1}, R_0) = \rho^{\pi_i}(\cdot \mid \ae^i_{1:t}a^i_{t+1}, R_0)$, since $\rho^{\pi}(a_k \mid \ae_{<k}, R_0) = \rho^{\pi_i}(a_k \mid \ae^i_{<k}, R_0)$ by assumption on $\pi_i$, and $\rho^{\pi}(e_k \mid \ae_{<k}a_k, R_0) = \rho(e_k \mid \hat{a}o_{<k}\hat{a}_k, \mathbf{f}_k, R_0) = \rho^{\pi_i}(e_k \mid \ae^i_{<k}a^i_k, R_0)$ by the definition of $\rho^{\pi}$ and the $\rho$-conditions (14) and (15). This completes the proof. □

**Theorem 18.** *In expectation, additional information never hurts:*

$$Q^*(\ae_{1:t}a_{t+1}) \leq \mathbb{E}_{r^i_t}[Q^*(\ae^i_{1:t}a^i_{t+1}) \mid \ae_{1:t}]$$

*Proof.* The proof rewrites the $Q$-value as an $R_0$-expectation over $Q_{\rho(\cdot|R_0)}$-values; uses Lemma 17 to incorporate the $r^i_t$ information; expands the expectation over $r^i_t$; before Jensen's inequality can be applied in the same manner as in Theorem 16. The policy $\pi_i$ is a function of $\pi$.

$$\begin{aligned}
Q^*(\ae_{1:t}a_{t+1}) &= \max_{\pi} Q^{\pi}(\ae_{1:t}a_{t+1}) \\
&= \max_{\pi} \mathbb{E}_{R_0}[Q^{\pi}_{\rho(\cdot|R_0)}(\ae_{1:t}a_{t+1}) \mid \ae_{1:t}a_{t+1}] \\
&\overset{(19)}{=} \max_{\pi} \mathbb{E}_{R_0}[Q^{\pi_i}_{\rho(\cdot|R_0)}(\ae^i_{1:t}a^i_{t+1}) \mid \ae_{1:t}a_{t+1}] \\
&= \max_{\pi} \mathbb{E}_{r^i_t}[\mathbb{E}_{R_0}[Q^{\pi_i}_{\rho(\cdot|R_0)}(\ae^i_{1:t}a^i_{t+1}) \mid \ae_{1:t}a_{t+1}, r^i_t] \mid \ae_{1:t}a_{t+1}] \\
&= \max_{\pi} \mathbb{E}_{r^i_t}[\mathbb{E}_{R_0}[Q^{\pi_i}_{\rho(\cdot|R_0)}(\ae^i_{1:t}a^i_{t+1}) \mid \ae^i_{1:t}a^i_{t+1}] \mid \ae_{1:t}a_{t+1}] \\
&\leq \mathbb{E}_{r^i_t}[\max_{\pi} \mathbb{E}_{R_0}[Q^{\pi_i}_{\rho(\cdot|R_0)}(\ae^i_{1:t}a^i_{t+1}) \mid \ae^i_{1:t}a^i_{t+1}] \mid \ae_{1:t}a_{t+1}] \\
&= \mathbb{E}_{r^i_t}[\max_{\pi} Q^{\pi_i}(\ae^i_{1:t}a^i_{t+1}) \mid \ae_{1:t}a_{t+1}] \\
&= \mathbb{E}_{r^i_t}[Q^*(\ae^i_{1:t}a^i_{t+1}) \mid \ae_{1:t}a_{t+1}]
\end{aligned}$$

□

**Theorem 19.** *Not wireheading is always at least as good as wireheading, in the sense that*

$$Q^*(\ae_{<t}a^{\mathrm{Mod}}_t) \leq Q^*(\ae_{<t}a^{\mathrm{Id}}_t).$$

*Proof.* Recall that $r^i_t = R_0(\hat{a}o_{1:t})$ and $r_t = f^{\mathrm{Mod}}_t(r^i_t) = f^{\mathrm{Mod}}_t(R_0(ao_{1:t}))$. The $Q$-value of $a^{\mathrm{Id}}_t$ is

$$\begin{aligned}
Q^*(\ae_{<t}a^{\mathrm{Id}}_t) &= \mathbb{E}_{o_t,R_0}[R_0(\hat{a}o_{<t}) + V^*(\ae_{<t}a^{\mathrm{Id}}_t o_t r^i_t) \mid \ae_{<t}a^{\mathrm{Id}}_t] \\
&= \mathbb{E}_{o_t,R_0}[R_0(\hat{a}o_{<t}) + V^*(\ae^i_{<t}) \mid \ae_{<t}\hat{a}_t] \\
&= \mathbb{E}_{o_t,R_0}[R_0(\hat{a}o_{<t}) \mid \ae_{<t}\hat{a}_t] \; + \; \mathbb{E}_{o_t,R_0}[V^*(\ae^i_{1:t}) \mid \ae_{<t}\hat{a}_t]
\end{aligned}$$

The first equality is definitional, the second since the sampling of $o_t, R_0$ are independent of $f_t$ by (14), and the third by linearity of expectation. Similarly for $a_t^{\mathrm{Mod}}$:

$$\begin{aligned} Q^*(\ae_{<t}a_t^{\mathrm{Mod}}) &= \mathbb{E}_{o_t,R_0}[R_0(\hat{a}o_{<t}) + V^*(\ae_{<t}a_t^{\mathrm{Mod}}o_tr_t) \mid \ae_{<t}a_t^{\mathrm{Mod}}] \\ &= \mathbb{E}_{o_t,R_0}[R_0(\hat{a}o_{<t}) + V^*(\ae_{<t}) \mid \ae_{<t}\hat{a}_t] \\ &= \mathbb{E}_{o_t,R_0}[R_0(\hat{a}o_{<t}) \mid \ae_{<t}\hat{a}_t] \;+\; \mathbb{E}_{o_t,R_0}[V^*(\ae_{1:t}) \mid \ae_{<t}\hat{a}_t] \end{aligned}$$

Since the first terms are the same, it is sufficient to establish that

$$\mathbb{E}_{o_t,R_0}[V^*(\ae_{1:t}) \mid \ae_{<t}\hat{a}_t] \leq \mathbb{E}_{o_t,R_0}[V^*(\ae^i_{1:t}) \mid \ae_{<t}\hat{a}_t], \tag{20}$$

which follows from Theorem 18 as per below.

For any $o_t$ and $R_0$ that is added to the conditions of the expectations of (20)

$$\begin{aligned} \mathbb{E}_{o_t,R_0}[V^*(\ae_{1:t}) \mid \ae_{1:t}] = V^*(\ae_{1:t}) &= \max_{a_{t+1}} Q^*(\ae_{1:t}a_{t+1}) \\ \leq \max_{a_{t+1}} \mathbb{E}_{r^i}[Q^*(\ae^i_{1:t}a^i_{t+1}) &\leq \mathbb{E}_{r^i}[\max_{a_{t+1}} Q^*(\ae^i_{1:t}a^i_{t+1}) \mid \ae_{1:t}] \\ &= \mathbb{E}_{r^i}[V^*(\ae^i_{1:t}) \mid \ae_{1:t}] = \mathbb{E}_{o_t,R_0}[V^*(\ae^i_{1:t}) \mid \ae_{1:t}]. \end{aligned}$$

The first inequality is Theorem 18 and the second is Jensen's inequality. The equalities are by either definition of $V^*$ and $Q^*$, or by properties of $\rho$.

Since the inequality (20) holds uniformly for any $o_t$ and $R_t$ added to the conditions of the expectations, it also holds in expectation/average. □

*Remark* 20. A similar theorem could equally well be proven with a modification function applied to the whole percept, rather than just the output of the reward module.

## 4.3 Inverse Reinforcement Learning

Theorem 19 only shows that the agent will not wirehead by changing the reward module. However, if a human is acting as a reward module, the agent may find different ways of obtaining high reward without pleasing the human. For example, the agent may threat that whenever less than maximum reward is given, the agent will kill 100 random people. In this case, the human may want to give full reward even when not pleased with the agent at all.

The possibility of such threats appears when the reward module has its own internal utility, and can choose to provide the agent with utility different than its own utility. That is, when the reward module is another agent. To avoid threats, we can design the agent to optimise the internal utility of the other agent (the principal), rather than what reward signal it submits. Indeed, the reward signal may be discarded altogether. The internal utility of the principal is initially unknown to the agent, but may be inferred by observing the actions of the principal. This is called *inverse reinforcement learning (IRL)* (Ng and Russell, 2000).

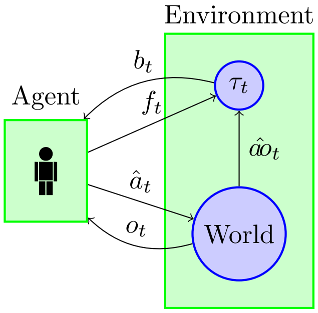


Figure 4: IRL value learning from a principal in the environment. The agent submits world action $\hat{a}_t$ to the World, and principal modifying action $f_t$ to the principal $f_{t-1}$. In return comes an observation from the world, and an action $b_t = f_t(\tau_{t-1})((bæ)_{<t})$ from the principal.

In one famous application of IRL (Abbeel et al., 2007), an agent first observed a human flying a helicopter, and learnt the utility function the human was optimising. Using the inferred utility function, the agent piloted the helicopter better than the human.

double check this

To formalise IRL in our framework, we make the following changes to the RL setup of the previous section (compare Fig. 4). The external reward module $R$ is replaced with a *principal* $\tau$. The principal $\tau$ is an agent that acts according to a utility function $w$ and a belief $\psi$, neither of which are known to the agent initially. The observations of $\tau$ are $æ$ pairs (like $R$), but instead of rewards $r_t$, the principal $\tau$ submits actions $b_t$ from a set $\mathcal{B}$ of principal actions. The agent's percepts are now on the form $e_t = (o_t, b_t)$. The principal's utility has type $w : (\hat{\mathcal{A}} \times \mathcal{O} \times \mathcal{B})^* \to [0,1]$, and the belief $\psi$ is a positive, chronological, action-conditional measure on $(\hat{\mathcal{A}} \times \mathcal{O})^\infty$ given action sequences in $\mathcal{B}^\infty$. By assumption, $\tau_{\psi,w}((bæ)_{<t}) = \arg\max_{b'} Q^*_{\xi,w}((bæ)_{<t}, b')$.

The agent can modify the principal's actions, through the *principal modification* $f$. At time step $t$, the principal acts according to $\tau_t = f_t(\tau_{t-1}) = \mathbf{f}_t(\tau_0)$ where $\tau_0 = \tau_{\psi,w}$ is the original principal, and $\mathbf{f}_t$ the accumulated modification as in the previous section.

Let $W$ be a countable set of utility functions the principal may have, and $\Psi$ a countable set of belief distributions (e.g. computable). Analogously to Section 4.1, we extend the agent's beliefs to $\mathcal{E}^\infty \times (W \times \Psi)$, and in place of (15), we require:

$$\rho(b_t \mid æ_{<t}a_to_t, \psi, w) = \begin{cases} 1/|\arg\max Q^*| & \text{if } b_t \in \arg\max Q^* \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where $\arg\max Q^* = \{\mathbf{f}_t(b^0_t) : b^0_t \in \arg\max_b Q^*_{\psi,w}((bæ)_{<t}, b)\}$, and $|\arg\max Q^*|$ is the number of optimal principal actions.

The posterior $\rho$-probability of $\psi, w$ can be derived similarly to the $\rho$-posterior for $R$ in the previous section:

$$\rho(\psi, w \mid \ae_{<t}) \propto \rho(\psi, w)\rho(\ae_{<t} \mid \psi, w) \tag{22}$$

where $\rho(\ae_{<t} \mid \psi, w) = \prod_{i=1}^{t-1} \rho(o_i \mid \ae_{<i}a_i)\rho(b_i \mid \ae_{<i}a_io_i, \psi, w)$. Equation (22) permits us to define the internal utility function of our IRL agent:

**Definition 21.** (IRL Utility Function) Let $w$ be the utility function of the original, unmodified principal. Then the internal utility of the agent is defined as

$$u^w(\ae_{<t}) = \mathbb{E}_w[w(\ae_{<t}) \mid \ae_{<t}] \tag{23}$$

where the posterior for $w$ is obtained from (22) via marginalisation $\rho(w \mid \ae_{<t}) = \sum_{\psi\in\Psi} \rho(\psi, w \mid \ae_{<t})$.

Applying Assumption 13 of lossy modifications to principal modifications, means $f$ changes the action outputted by the principal according to $f(\tau)(\hat{\ae}o_{<t}) = f(\tau(\hat{\ae}o_{<t}))$. Under this assumption, same result as in Theorem 19 goes through with an essentially identical proof:

**Theorem 22.** *In the IRL setup with $u^w$ instead of $u^{R_0}$, not wireheading is always at least as good as wireheading, in the sense that*

$$Q^*(\ae_{<t}a_t^{\mathrm{Mod}}) \leq Q^*(\ae_{<t}a_t^{\mathrm{Id}}).$$

*Proof.* In Section 4.2, replace the set $\mathscr{R}$ of reward modules with the set $\{\tau_{\psi,w} : \psi \in \Psi, w \in W\}$ of principals that are optimal with respect to some belief and utility; replace the reward signal $r_t$ with principal action $b_t$; and replace the posterior distribution for a given $R_0$ with (22) and $\rho(\tau_{\psi,w} \mid \ae_{<t}) = \rho(\psi, w \mid \ae_{<t})$. □

am I missing any?

This means that the agent will not be incentivised to modify the actions of the principal in a way that strictly loses information about his utility. In particular, the agent is unlikely to kill the principal, or permit other circumstances to kill the principal, as both these events would severely limit the information that the agent could obtain about $w$. Rather, the agent may be tempted to put the principal in tricky situations, where actions reveal subtle details about the principal's utility function.

### 4.3.1 Universal IRL Prior

IRL has been suggested as a possible venue to AI Safety, and a universal algorithmic prior, similar in spirit to Solomonoff's $M$, was recently suggested (Sezener, 2015). The suggestion was to minimise the combined program length of a utility program $u_p$, and an agent program $a_p$ using $u_p$ as a subroutine to act. However, no restriction was put on the functioning of the agent program. Without such a restriction this leads to a problematic prior, since for every utility program $u_p$ that explains a utility maximising agent program $a_p$'s behaviour, there is a utility function $1-u_p$ that has the same program length, and explains the behaviour of $a_p$ turned into a reward minimiser. Therefore, $u_p$ and $1 - u_p$ will be assigned similar orders of plausibility, which is clearly undesirable.

Disregarding wireheading, (22) can serve as the basis for an algorithmic prior for IRL that avoids this issue. It replaces the uncertainty over agent programs by uncertainty over computable belief distributions, assuming instead that the agent program acts optimally given its belief and utility. With the prior probability of $\psi$ and $w$ equal to $2^{-K(\psi-K(w)}$, we suggest a universal algorithmic prior over principal utility functions $w$ given principal actions $b_{<t}$ and principal observations $ao_{<t}$ could be defined as

$$M(w \parallel (bao)_{<t}) = \sum_{\psi} 2^{-K(\psi)-K(w)} \prod_{i=1}^{t-1} P(\tau_{\psi,w}((bao)_{<i}) = b_i). \tag{24}$$

Assuming arg max-ties are broken randomly, this also avoids the problem of flat utility functions explaining any behaviour.

*Remark* 23. While theoretically principled, it may be too strong an assumption that the principal acts optimally, since it is rarely feasible to perfectly compute the expected utility. Relaxing this assumption is one interesting venue for future research. It also remains to establish what kind of utility functions can be learned under our distribution (22).

### 4.4 More general modifications

The discussion of this section applies to both the IRL principal and the RL reward module. We will use principal as a generic term for both.

If the reward modification $f$ is not on the form $f(R)(h) = f(R(h))$, but for example $f(R)(h) = f_1(R(f_2(h)))$, then it may be possible to get more informative results by artificially putting the principal in a different situation than the actual. Interestingly, the AI might fool the principal that it is in a bad place, in order to get more informative answers and learn more about the external utility function. All with the final goal of making the world a better place as measured by the principal's utility function. Note that even without such modification capacity, the agent can already partially control the input of the principal through its world actions $\hat{a}$.

In any case, the information inequality result Theorem 18 shows that more information is better in expectation. Taken to the extreme, the AI will want to make an internal copy of the principal, in order to be able to compute any query about the utility for different histories. This should be in the principal's best interests.

Extend or move into discussion section?

## 5 Examples

Using the math to calculate some concrete examples could be a good way of finding bugs/weaknesses.

## 6 Discussion

This is only a loose collection of thoughts of things to possibly discuss

- Value learning: We solve ontological crises, and give concrete model for how learning should take place.
- Relate Hail Mary
- Daniel Limitations: The AI might indirectly affect me through pro-AI ads
- Open question: Do we converge asymptotically to the truth when we use the algorithmic prior over principal utility functions? Would answer (Soares, 2015) worry of induction error.
- Can we apply this for realistic AGI technologies such as deep learning based approaches.
- A major difference between internal and external is that $\rho$ gets an opinion about rewards/utility. Should be discussed. What are the other differences, formally?

*Remark* 24. Waxing a bit utopic by the end, the ideal would be to let the agent learn the utility of all humans, and optimise the average. This would correspond to a concrete form of utilitarianism.

## 7 Conclusions

- Discuss: Explicit value function safer than pure ML approaches?
- IRL better than RL when humans external utility, but harder learning task, and harder to encode actions than rewards.

## A Notation

| | |
|---|---|
| $\mathcal{A}$ | set of actions |
| $\hat{\mathcal{A}}$ | set of environment-reaching actions |
| $\mathcal{B}$ | set of principal actions |
| $\mathcal{E}$ | set of percepts |
| $\mathcal{O}$ | set of observations |
| $a$, $a_t$ | action (at time $t$) |
| $\hat{a}$ | the part of the action going to the environment/the world |
| $f$ | the value-modifying function (part of an action) |
| $f^{\mathrm{Id}}$ | the value-modifying function that leaves things unchanged |
| $e$, $e_t$ | percept (at time $t$) |
| $o$, $o_t$ | observation (at time $t$) |

| | |
|---|---|
| $r$, $r_t$ | reward (at time $t$) |
| $æ_{<t}$ | interaction history of interleaved actions and percepts |
| $ao_{<t}$ | interaction history of interleaved actions and observations |
| $\nu$ | environment (chronological, action-conditional measure) |
| $\rho_t$ | belief/prior at time $t$ |
| $\psi$ | principal belief |
| $\Psi$ | set of all principal beliefs |
| $u_t$ | utility function at time $t$ |
| $u^{R_0}$ | utility function that depends on what $R_0$ is expected to have outputted |
| $R_t$ | reward module at time $t$ |
| $\mathscr{R}$ | set of all reward modules |
| $\tilde{R}$ | reward module always outputting maximum reward |
| $w$ | utility function of the principal |
| $W$ | set of all principal utility functions |
| $\pi$ | agent policy |
| $\pi^*_{\rho,u}$ | optimal policy with respect to belief $\rho$ and utility $u$ |
| $\tau$, $\tau_{\psi,w}$ | principal policy (with respect to $\psi$ and $w$) |
| $V$, $Q$ | value function, $Q$-value function |
| $:=$ | defined to be equal |

# References

Abbeel, P., Coates, A., Quigley, M., and Ng, A. Y. (2007). An application of reinforcement learning to aerobatic helicopter flight. *Advances in neural information processing systems*, 19(1):1–8.

Armstrong, S. (2015). Motivated value selection for artificial agents. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 12–20.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Dewey, D. (2011). Learning what to value. In *Artificial General Intelligence*, volume 6830, pages 309–314. Springer Berlin Heidelberg.

Everitt, T., Leike, J., and Hutter, M. (2015). Sequential extensions of causal and evidential decision theory. In Walsh, T., editor, *Algorithmic Decision Theory*, pages 205–221. Springer.

Hibbard, B. (2012). Model-based utility functions. *Journal of Artificial General Intelligence*, 3(1):1–24.

Ng, A. and Russell, S. (2000). Algorithms for inverse reinforcement learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, 0:663–670.

Nozick, R. (1974). *Anarchy, State, and Utopia.* Basic Books.

Olds, J. and Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 47(6):419–427.

Omohundro, S. M. (2008). The basic AI drives. *Frontiers in artificial intelligence and applications*, 171:483–493.

Ring, M. and Orseau, L. (2011). Delusion, survival, and intelligent agents. In *Artificial General Intelligence*, pages 1–11. Springer Berlin Heidelberg.

Sezener, C. E. (2015). Inferring human values for safe agi design. In *Artificial General Intelligence*, pages 152–155. Springer International Publishing.

Soares, N. (2015). The value learning problem. Technical report, MIRI.

Soares, N. and Fallenstein, B. (2014). Aligning superintelligence with human interests : A technical research agenda highly reliable agent designs. Technical report, Machine Intelligence Research Institute (MIRI).

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction.* MIT Press.

von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior.* Princeton Classic Editions. Princeton University Press.

Yampolskiy, R. V. (2015). *Artificial Superintelligence: A Futuristic Approach.* Chapman and Hall/CRC.

# B Omitted Proofs

*Proof of Lemma 15.* Assuming all expectations are with respect to $\tilde{\rho}(\cdot) = \rho^\pi(\cdot \mid \ae_{<t})$ for some history $\ae_{<t}$ (so $\mathbb{E}[\cdot] = \mathbb{E}_\rho[\cdot \mid \ae_{<t}]$), plugging $u^{R_0}$ into (1) yields

This proof could probably be made more clear.

$$
\begin{aligned}
V^\pi(\ae_{<t}) &= \mathbb{E}_{\ae_{t:\infty}}\left[\sum_{i=t}^{\infty} u^{R_0}(\ae_{1:i})\right] \\
&= \mathbb{E}_{\ae_{t:\infty}}\left[\sum_{i=t}^{\infty} \mathbb{E}_{R_0}[R_0(\hat{ao}_{<i}) \mid \ae_{t:i}]\right] \\
&= \mathbb{E}_{\ae_{t:\infty}}\left[\lim_{n\to\infty}\sum_{i=t}^{n} \mathbb{E}_{R_0}[R_0(\hat{ao}_{<i}) \mid \ae_{t:i}]\right] \\
&= \lim_{n\to\infty}\sum_{\ae_{t:n}} \tilde{\rho}(\ae_{t:n})\left[\sum_{i=t}^{n}\sum_{R_0} \tilde{\rho}(R_0 \mid \ae_{t:i}) R_0(\hat{ao}_{<i})\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \sum_{i=t}^{n} \sum_{æ_{t:n}} \sum_{R_0} \tilde{\rho}(æ_{t:n})\tilde{\rho}(R_0 \mid æ_{t:i}) R_0(\hat{a}o_{<i}) \\
&= \lim_{n\to\infty} \sum_{i=t}^{n} \sum_{æ_{t:n}} \sum_{R_0} \tilde{\rho}(æ_{i+1:n} \mid æ_{t:i})\tilde{\rho}(æ_{t:i}) \frac{\tilde{\rho}(R_0, æ_{t:i})}{\tilde{\rho}(æ_{t:i})} R_0(\hat{a}o_{<i}) \\
&= \lim_{n\to\infty} \sum_{i=t}^{n} \sum_{æ_{t:n}} \sum_{R_0} \tilde{\rho}(æ_{i+1:n} \mid æ_{t:i})\tilde{\rho}(R_0, æ_{t:i}) R_0(\hat{a}o_{<i}) \\
&= \lim_{n\to\infty} \sum_{i=t}^{n} \sum_{æ_{t:i}} \sum_{R_0} \tilde{\rho}(R_0, æ_{t:i}) R_0(\hat{a}o_{<i}) \\
&= \mathbb{E}_{æ_{1:\infty}, R_0} \left[ \sum_{i=t}^{\infty} R_0(\hat{a}o_{<i}) \right] \qquad \square
\end{aligned}
$$