Performance of Bounded-Rational Agents With the Ability to Self-Modify*

Jakub Tětek[1†], **Marek Sklenka**[2], **Tomáš Gavenčiak**[3]
[1] BARC, University of Copenhagen j.tetek@gmail.com
[2] University of Oxford sklenka.marek@gmail.com
[3] Independent researcher gavento@ucw.cz

Abstract

Self-modification of agents embedded in complex environments is hard to avoid, whether it happens via direct means (e.g. own code modification) or indirectly (e.g. influencing the operator, exploiting bugs or the environment). It has been argued that intelligent agents have an incentive to avoid modifying their utility function so that their future instances work towards the same goals.

Everitt et al. (2016) formally show that providing an option to self-modify is harmless for perfectly rational agents. We show that this result is no longer true for agents with bounded rationality. In such agents, self-modification may cause exponential deterioration in performance and gradual misalignment of a previously aligned agent. We investigate how the size of this effect depends on the type and magnitude of imperfections in the agent's rationality (1-4 below). We also discuss model assumptions and the wider problem and framing space.

We examine four ways in which an agent can be bounded-rational: it either *(1)* doesn't always choose the optimal action, *(2)* is not perfectly aligned with human values, *(3)* has an inaccurate model of the environment, or *(4)* uses the wrong temporal discounting factor. We show that while in the cases *(2)-(4)* the misalignment caused by the agent's imperfection does not increase over time, with *(1)* the misalignment may grow exponentially.

1 Introduction

We face the prospect of creating superhuman (or otherwise very powerful) AI systems in the future where those systems hold significant power in the real world (Bostrom 2014; Russell 2019). Building up theoretical foundations for the study and design of such systems gives us a better chance to align them with our long-term interests. In this line of work, we study agent-like systems, i.e. systems optimizing their actions to maximize a certain utility function – the framework behind the current state-of-the-art reinforcement learning systems and one of the major proposed models for future AI systems[1].

If strong AI systems with the ability to act in the real world are ever deployed[2], it is very likely that they will have some means of deliberately manipulating their own implementation, either directly or indirectly (e.g. via manipulating the human controller, influencing the development of a future AI, exploiting their own bugs or physical limitations of the hardware, etc). While the extent of those means is unknown, even weak indirect means could be extensively exploited with sufficient knowledge, compute, modelling capabilities and time.

Omohundro (2008) argues that every intelligent system has a fundamental drive for goal preservation, because when the future instance of the same agent strives towards the same goal, it is more likely that the goal will be achieved. Therefore, Ohomundro argues, a rational agent should never modify into an agent optimizing different goals.

Everitt et al. (2016) examine this question formally and arrive at the same conclusion: that the agent preserves its goals in time (as long as the agent's planning algorithm anticipates the consequences of self-modifications and uses the current utility function to evaluate different futures).[3] However, Everitt's analysis assumes that the agent is a perfect utility maximizer (i.e. it always takes the action with the greatest expected utility), and has perfect knowledge of the environment. These assumptions are probably unattainable in any complex environment.

To address this, we present a theoretical analysis of a self-modifying agent with imperfect optimization ability and incomplete knowledge. We model the agent in the standard

*Supported by Grant Number 16582, Basic Algorithms Research Copenhagen (BARC), from the VILLUM Foundation.

†This author was supported by the Bakala Foundation Scholarship

[1] Other major models include e.g. comprehensive systems of services (Drexler 2019) and ”Oracle AI” or ”Tool AI” (Armstrong, Sandberg, and Bostrom 2012). However, there are concerns and ongoing research into the emergence of agency in these systems (Omohundro 2008; Miller, Yampolskiy, and Häggström 2020).

[2] Proposals to prevent this include e.g. boxing (Bostrom 2014) but as e.g. Yampolskiy (2012) argues, this may be difficult or impractical.

[3] Everitt et al. (2016)'s results hold independent of the length of the time horizon or temporal discounting (by simple utility scaling).

cybernetic model where the agent can be bounded-rational in two different ways. Either the agent makes suboptimal decisions (is a bounded-optimization agent) or has inaccurate knowledge. We conclude that imperfect optimization can lead to exponential deterioration of alignment through self-modification, as opposed to bounded knowledge, which does not result in future misalignment. An informal summary of the results is presented below.

Finally, we explicitly list and discuss the underlying assumptions that motivate the theoretical problem and analysis. In addition to clearly specifying the *scope of conclusions*, the explicit problem assumptions can be used as a rough axis to map the space of viable research questions in the area; see Sections 2 and 6.

1.1 Summary of our results

The result of Everitt et al. (2016) could be loosely interpreted to imply that agents with close to perfect rationality would either prefer not to self-modify, or would self-modify and only lose a negligible target value.

We show that when we relax the assumption of perfect rationality, their result no longer applies. The bounded-rational agent may prefer to self-modify given the option and in doing so, become less aligned and lose a significant part of the attainable value according to its original goals.

We use the difference between the attainable and attained expected future value at an (arbitrarily chosen) future time point as a proxy for the degree of the agent's misalignment at that time. Specifically, for a future time t, we consider the value attainable from time t (after the agent already ran and self-modified for t time units), and we estimate the loss of value f^t relative to the non-modified agent in the same environment state. Note that f^t is not pre-discounted by the previous t steps. See Section 3 for formal definitions and Section 2 for motivation and discussion.

We consider four types of deviation from perfect rationality, see Section 4 for formal definitions.

- ϵ-*optimizers* make suboptimal decisions.
- ϵ-*misaligned agents* have inaccurate knowledge of the human utility function.
- ϵ-*ignorant agents* have inaccurate knowledge of the environment.
- ϵ-*impatient agents* have inaccurate knowledge of the correct temporal discount function.

Note that for the sake of simplicity, we use a very simple model of bounded rationality where the errors are simply bounded by the error parameters $\epsilon_\bullet$; this has to be taken into account when interpreting the results. However, we suspect that the asymptotic dependence of value loss on the size of errors and time would be similar for a range of natural, realistic models of bounded rationality.

Informal result statements

Self-modifying ϵ-*optimizers* may deteriorate in future alignment and performance exponentially over time, losing exponential amount of utility compared to ϵ-optimizers that do not self-modify. We show upper and tight lower bounds (by a constant) on the worst-case value loss in Theorem 7. As we decrease γ (increase discounting), the rate at which the agent's performance deteriorates increases and the possibility of self-modification becomes a more serious problem.

Our analysis of bounded-optimization agents is a generalization of Theorem 16 from Everitt et al. (2016) in the sense that their result can be easily recovered by a basic measure-theoretic argument.

Self-modifying ϵ_u-*misaligned,* ϵ_ρ-*ignorant, or* ϵ_γ-*impatient perfect optimizers* can only lose the same value as non-self-modifying agents with the same irrationality bounds. This also holds for any combination of the three types of bounded knowledge. We give tight upper and lower bounds (up to a constant factor) for the worst-case performance. See Section 5.2 for details.

This implies that unlike bounded-optimization agents, the performance of perfect-optimization bounded-knowledge agents does not deteriorate in time. This is because bounded-knowledge agents continue to take optimal actions with respect to their almost correct knowledge and do not self-modify in a way that would worsen their performance in their view. Therefore, the possibility of self-modification seems less dangerous in the case of bounded-knowledge agents than in the case of bounded-optimization agents.

A *self-modifying agent with any combination of the four irrationality types* may lose value exponential in the time step t when the agent optimization error parameter $\epsilon_o > 0$. We again give tight (up to a constant factor) lower bounds on the worst-case performance of such agents. See Section 5.3 for details.

Our results do not imply that every such agent will actually perform this poorly but the prospect of exponential deterioration is worrying in the long-term, even if it happens at a much slower speed than suggested by our results. We focus on worst-case analysis because it tells us whether we can have formal guarantees of the agent's behaviour – a highly desirable property for powerful real-world autonomous systems, including a prospective AGI (artificial general intelligence) or otherwise strong AIs.

Overview of formal results. Here we summarize how much value the different types of bounded-rational agents may lose via misalignment. Note that the maximal attainable discounted value is at most $\frac{1}{1-\gamma}$ and the losses should be considered relative to that, or to the maximum attainable value in concrete scenarios. Otherwise, the values for different value of γ are incomparable. In all cases, the worst-case lower and upper bounds are tight up to a constant.

ϵ-*optimizer agents* – bounded optimization, after t steps of possible self-modification (Theorem 7)

$$f^t_{\text{opt}}(\epsilon, \gamma) = \min(\frac{\epsilon}{\gamma^{t-1}}, \frac{1}{1-\gamma})$$

ϵ-*misaligned agents* – inaccurate utility (Theorem 9)

$$f_{\text{util}}(\epsilon, \gamma) = \frac{2\epsilon}{1-\gamma}$$

ϵ-ignorant agents – inaccurate belief (Theorem 11)

$$f_{\text{bel}}(\epsilon, \gamma) = \frac{2}{1-\gamma} - \frac{2}{1-\gamma(1-\epsilon)}$$

ϵ-impatient agents – inaccurate discounting (Theorem 13) Here γ^* is the correct discount factor and γ is the agent's incorrect discount factor.

$$f_{\text{disc}}(\gamma, \gamma^*) \approx \frac{2\gamma^{*\,\frac{1}{\lg \gamma}} - 1}{1-\gamma^*}$$

2 Assumptions and rationale

Both the statement of the problem and its relevance to AI alignment rest on a set of assumptions listed below. While this list is non-exhaustive, we try to cover the main implicit and explicit choices in our framing, and the space of alternatives. This is largely in hope of eventually finding a better, more robust theoretical framework for solving agent self-modification within the context of AI alignment, but even further negative results in the space would inform our intuitions on what aspects of self-modification make the problem harder.

We propose consideration of various assumptions as a framework for thinking about prospective *realistic agent models that admit formal guarantees*. We invite further research and generalizations in this area, one high-level goal being to map a part of the space of agent models and assumptions that do or do not permit guarantees, eventually finding agent models that do come with meaningful guarantees. Further negative results would inform our intuitions on what aspects of the problems make it harder.

(i) *Bounded rationality model.* In the models of ϵ-bounded-rational agents defined in Section 4.1, ϵ is generally an upper bound on the size of the optimization or knowledge error. One interpretation of our results is that value drift can happen even if the error is bounded at every step. One could argue that a more realistic scenario would assume some distribution of the size of the errors, assuming larger errors less likely or less frequent; see discussion below and in Section 6.

(ii) *Unlimited self-modification ability.* We assume the agent is able to perform any self-modification at any time. This models the worst-case scenario when compared to a limited but still perfectly controlled self-modification. However, embedded (non-dualistic) agents in complex environments may chieve almost-unlimited self-modification from a limited ability, e.g. over a longer time span; see e.g. (Demski and Garrabrant 2019). We model the agent's self-modifications as orthogonal to actions in the environment.

(iii) *Modification-independence.* We assume that the agent's utility function does not explicitly reward or punish self-modifications. We also assume that self-modifications do not have any direct effect on the environment. This is captured by Definition 2.

(iv) *No corrigibility mechanisms.* We do not consider systems that would allow human operators to correct the system's goals, knowledge or behaviour. The problem of robust strong AI corrigibility is far from solved today and this paper can be read as a further argument for substantially more research in this direction.

(v) *Worst-case analysis and bound tightness.* We focus on worst-case performance guarantees in abstracted models rather than e.g. full distributional analysis, and we show that our worst-case bounds are attainable (up to constant factors) under certain agent behaviour. Note this approach may turn out as too pessimistic or even impossible in some settings (e.g. quantum physics).

(vi) *Bounded value attainable per time unit.* We assume the agent obtains instantaneous utility between 0 and 1 at each time step. This is not an arbitrary choice: A constant bound on instantaneous value can be normalized to this interval. Instantaneous values bounded by a function of time $U(t) < \mu^t$ can be pre-discounted when $\gamma\mu < 1$, and generally lead to infinite future values otherwise, which we disallow here to avoid foundational problems.

(vii) *Temporal value discounting.* We assume the agent employs some form of temporal value discounting. This could be motivated by technical or algorithmic limitations, increasing uncertainty about the future, or to avoid issues with incomparable infinite values of considered futures (see Bostrom (2011) for a discussion of infinite ethics). Discounting, however, contrasts with the long-termist view; see the discussion below.

(viii) *Exponential discounting.* Our model assumes the agent discounts future utility exponentially, a standard assumption in artificial intelligence and the only time-invariant discounting schema (Strotz 1955) leading to consistent preferences over time.

(ix) *Unbounded temporal horizons.* Our analysis focuses on the long-term behaviour of the agent, in particular stability and performance from the perspective of future stakeholders (sharing the original utility function). Note that our results also to some extent apply to finite-horizon but long-running systems.

Temporal discounting contrasts with the long-termist view: Why not model non-discounted future utility directly? Noting the motivations we mention in (vii), we agree that models of future value aggregation other than discounting would be generally better suited for long-term objectives. However, this seems to be a difficult task, as such models are neither well developed nor currently used in open-ended AI algorithms (with the obvious exception of a finite time horizon, which we propose to explore in Section 6).

We therefore propose a direct interpretation of our results: *Assuming we implement agents that are ϵ-optimizers with discounting, they may become exponentially less aligned over time. This is not the case with perfect optimizers with imperfect knowledge and discounting.*

(x) *Dualistic setting.* We assume a dualistic agent and allow self-modification through special actions. This allows us to formally model one aspect of embedded agency – at least until there are sufficient theoretical foundations of embedded agency.

Note that in the embedded (non-dualistic) agent setting, it is not formally clear – or possibly even definable – what constitutes a self-modification, since there is no clear conceptual boundary between the agent and the environment, as discussed by Demski and Garrabrant (2019).

Assumption categories and the problem space. Each assumption identifies a subspace of research questions we would obtain by varying the relevant choices. These subspaces vary from very technical (e.g. concrete rationality model) to foundational (e.g. finite values and dualistic agent models). Along this axis, the assumptions and choices point to different kinds of prospective problems; we briefly describe three such categories and their prospects. See Section 6 for concrete proposals of future work.

Technical choices: A concrete model of bounded rationality, unlimited self-modification model, and modification-independence. These are likely important for short and medium time-frames, where even eventually-diverging guarantees are useful.

We believe that many models within some realistic and sufficiently strong model classes would lead to qualitatively equivalent results in long time horizons; e.g. the agent divergence would be asymptotically exponential without external corrigibility, embedded agents in sufficiently complex environments would be able to self-modify arbitrarily over a long time (see discussion above) etc. However, these intuitions call for further verification.

Problem components: No corrigibility mechanisms, unbounded time horizon, time-invariant temporal discounting, focus on the worst-case guarantees. For those, there are interesting alternatives that may yield more optimistic results. In particular, it would be valuable to explore formal models of corrigibility, perform a full probabilistic analysis of agent development, and develop long-term non-discounted finite-time settings.

Foundational assumptions: Dualistic agent model and finite value of the future. Those are a standard in the area, but alternative settings may open up important and fruitful model classes and technical choices that capture currently pre-paradigmatic aspects (e.g. theory of embedded agency and non-dualistic agents).

3 Preliminaries

In this section, we explain our model of a self-modifying agent, which is borrowed from Everitt et al. (2016). We will extend this model to include bounded rationality in Section 4.

We use a modified version of the *standard cybernetic model*. In this model, an agent interacts with the environment in discrete time steps. At each time step t, the agent performs an *action* a_t from a finite set $\mathcal{A}$ and the environment responds with a *perception* e_t from a finite set $\mathcal{E}$. An *action-perception pair* $æ_t$ is an action concatenated with a perception. A *history* is a sequence of action-perception pairs $æ_1 æ_2 ... æ_t$. We will often abbreviate such sequences to $æ_{<t} = æ_1 ... æ_{t-1}$ or $æ_{n:m} = æ_n ... æ_m$. A *complete* history $æ_{1:\infty}$ is a history containing information about all the time steps.

An agent can be described by its policy π. The policy[4] $\pi\colon (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$ is used to determine the agent's next action from the history at time t. We consider (bounded-rational) utility maximizers, where the policy is (partially) determined by the instantaneous utility function u, belief ρ and discount factor γ. We sometimes use the notation $\kappa = (u, \rho, \gamma)$, where κ is called the agent's *knowledge*. The utility function $\tilde{u}\colon (\mathcal{A} \times \mathcal{E})^\infty \to \mathbb{R}$ describes how much the agent prefers the complete history $æ_{1:\infty}$ compared to other complete histories. We will assume that the total utility is a discounted sum of instantaneous utilities given by the instantaneous utility function $u\colon (\mathcal{A} \times \mathcal{E})^* \to [0, 1]$. Formally, $\tilde{u}(æ_{1:\infty}) = \sum_{t=1}^{\infty} \gamma^{t-1} u(æ_{\leq t})$. The discount factor γ describes how much the agent prefers immediate reward compared to the same reward at a later time. Smaller γ means heavier discounting of the future and stronger preference for immediate reward. Note that the maximum achievable utility is $\frac{1}{1-\gamma}$, which happens when $u(æ_t) = 1$ at each step. Also note that instantaneous utility depends not only on the latest perception but can also depend on all previous perceptions and actions.

In addition to all this, an agent has a belief $\rho\colon (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \to \Delta\bar{\mathcal{E}}$ where $\Delta\bar{\mathcal{E}}$ is the set of full-support probability distributions over $\mathcal{E}$. This is a function which maps any history ending with an action onto a probability distribution over the next perceptions. Intuitively speaking, the belief describes what the agent expects to see after it performs an action given a certain history. A belief together with a policy induce a measure on the set $(\mathcal{A} \times \mathcal{E})^*$ using $P(e_t \mid æ_{<t} a_t) = \rho(e_t \mid æ_{<t} a_t)$ and $P(a_t \mid æ_{<t}) = 1$ if $\pi(æ_{<t}) = a_t$ and 0 otherwise. Intuitively speaking, this probability measure captures probabilities assigned by the agent to possible futures.

Following the reinforcement learning literature, we define the *value function* $V\colon (\mathcal{A} \times \mathcal{E})^* \to \mathbb{R}$ as the expected future discounted utility:

$$V^\pi(æ_{<t}) = \mathbb{E}[\sum_{t'=t}^{\infty} \gamma^{t'-t} u(æ_{<t'})\,]$$

The expectation value on the right is calculated with respect to belief ρ and assuming the agent will follow the policy π. Intuitively, the value function describes how promising the future seems. When the value $V^\pi(æ_{<t})$ of a history is high, it means we can expect an agent with policy π to collect a lot of utility in the future starting from this history. Note that when calculating V-values, instantaneous utilities are multiplied by $\gamma^{t'-t}$ rather than $\gamma^{t'}$. This means that V-values can remain high throughout the whole history and are not affected by discounting. We define the Q-value of an action as the expected future discounted utility after taking that ac-

[4] For a set S, S^* denotes the set of finite sequences of elements from S

tion:

$$Q^{\pi}(\text{\ae}_{<t}a_t) = \mathbb{E}[u(\text{\ae}_{1:t}) + V^{\pi}(\text{\ae}_{1:t})]$$

where the expectation is over the next perception drawn from the belief (note that belief is a probability distribution)

The Q-value measures how good an action is given that the agent will later follow policy π. A policy π^* is an *optimal policy* when $V^{\pi^*}(\text{\ae}_{<t}) = \sup_{\pi} V^{\pi}(\text{\ae}_{<t})$ for all histories $\text{\ae}_{<t}$ (such a policy always exists, as shown in (Lattimore and Hutter 2014)).

3.1 Self-modification model

In this section, we extend the formalism above to include the possibility of self-modification. Since we are interested in the worst-case scenario, we assume the agent has unlimited self-modification ability. Worst-case results derived for such an agent will also hold for agents with limited ability to self-modify.

Definition 1. *A policy self-modification model is defined as a quadruple $(\breve{\mathcal{A}}, \mathcal{E}, \mathcal{P}, \iota)$ where $\breve{\mathcal{A}}$ is the set of world actions, $\mathcal{E}$ is the set of perceptions, $\mathcal{P}$ is a non-empty set of names and ι is a map from $\mathcal{P}$ to the set of all policies Π.*

At every time step, the agent chooses an action $a_t = (\breve{a}_t, p_{t+1})$ from the set $\mathcal{A} = (\breve{\mathcal{A}} \times \mathcal{P})$. The first part $\breve{a}_t$ describes what the agent "actually does in the world" while the second part chooses the policy $\pi_{t+1} = \iota(p_{t+1})$ for the next time step. We will also use the notation $a_t = (\breve{a}_t, \pi_{t+1})$, keeping in mind that only policies with names may be chosen. This new policy is used in the next step to pick the action $a_{t+1} = \pi_{t+1}(\text{\ae}_{1:t})$. Note that only policies with names can be chosen and that $\mathcal{P} = \Pi$ is not a possibility because it entails a contradiction: $|\Pi| = |(\breve{\mathcal{A}} \times \mathcal{E} \times \Pi)|^{|(\breve{\mathcal{A}} \times \mathcal{E} \times \Pi)^*|} > 2^{|\Pi|} > |\Pi|$. A history can now be written as:

$$\text{\ae}_{1:t} = a_1 e_1 a_2 e_2 ... a_t e_t = \breve{a}_1 \pi_2 e_1 \breve{a}_2 \pi_3 e_2 ... \breve{a}_t \pi_{t+1} e_t$$

The subscripts for policies are one time step ahead because the policy chosen at time t is used to pick an action at time $t+1$. The subscript denotes at which time step the policy is used. Policy π_t is used to choose the action $a_t = (\breve{a}_t, \pi_{t+1})$. No policy modification happens when $a_t = (\breve{a}_t, \pi_t)$.

In the previous section, we used these rules to calculate the probability of any finite history: $P(e_t \mid \text{\ae}_{<t}a_t) = \rho(e_t \mid \text{\ae}_{<t}a_t)$ and $P(a_t \mid \text{\ae}_{<t}) = 1$ if $\pi(\text{\ae}_{<t}) = a_t$ and 0 otherwise. However, the second rule doesn't take into consideration that the agent's policy is changing. Therefore, to account for self modification, we need to modify the second rule into "$P(a_t \mid \text{\ae}_{<t}) = 1$ if $\pi_t(\text{\ae}_{<t}) = a_t$ and zero otherwise". To evaluate the V and Q-functions for self-modifying agents, we need to use probabilities of complete histories calculated in this way.

Definition 2. *Let $\breve{\text{\ae}}_{1:t}$ denote the history $\text{\ae}_{1:t}$ with information about self-modification removed so that $\breve{\text{\ae}}_{1:t} = \breve{a}_1 e_1 \breve{a}_2 e_2 ... \breve{a}_t e_t$. A function $f \colon (\mathcal{A} \times \mathcal{E})^* \to$ (anything) is modification-independent if $\breve{\text{\ae}}_{1:t} = \breve{\text{\ae}}'_{1:t}$ implies that $f(\text{\ae}_{1:t}) = f(\text{\ae}'_{1:t})$.*

Modification-independence assumption: In the rest of the paper, we will assume that the agent's belief and utility function as well as the correct belief are modification-independent.

4 Definitions of bounded-rational agents

We now extend the model from Everitt et al. (2016) by defining two types of bounded-rational agents which we will be using throughout the paper: bounded-optimization agents (described in Section 4.1) and bounded-knowledge agents (described in Section 5.3). Bounded-knowledge agents can be subdivided further into misaligned, ignorant and impatient agents.

4.1 Bounded-optimization agents

We introduce the notion of *ϵ-optimizers*. Intuitively speaking, the expected future discounted utility gained by an ϵ-optimizer is no more than ϵ lower than the optimal one in any situation they could get into (that is, for any history).

Definition 3. *We say that agent A is an ϵ-optimizer for history $\text{\ae}_{<t}$ if it holds that*

$$Q(\text{\ae}_{<t}\pi(\text{\ae}_{<t})) \geq \sup_{\pi'} Q(\text{\ae}_{<t}\pi'(\text{\ae}_{<t})) - \epsilon$$

When the utility function, belief and discount factor is obvious from the context (or unimportant), we also speak of policy being ϵ-optimizing (with respect to the utility function and belief), meaning that the corresponding agent is an ϵ-optimizer.

4.2 Bounded-knowledge agents

We consider agents with inaccurate knowledge of the correct utility function (Definition 4), inaccurate knowledge of the world (Definitions 5 and 6), and inaccurate knowledge of the correct discount factor (how much future reward is worth compared to reward in the present).

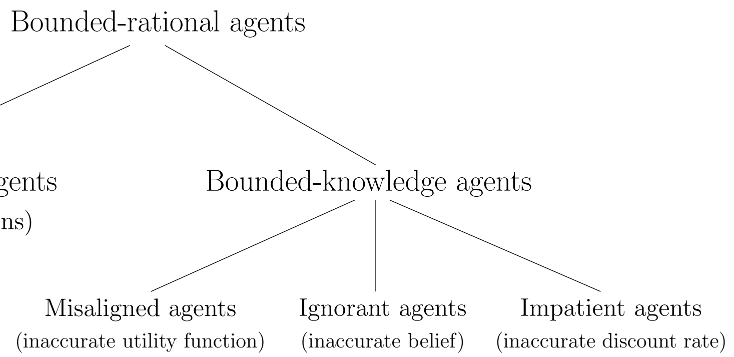
Misaligned agents We define ϵ-misaligned agents as agents whose utility function u has absolute error ϵ with respect to the correct utility function u^*.

Definition 4. *We say that the instantaneous utility function u has absolute error ϵ with respect to the correct utility function u^* if*

$$\sup_{t \in \mathbb{N}, \text{\ae}_{<t}} |u(\text{\ae}_{<t}) - u^*(\text{\ae}_{<t})| = \epsilon$$

Ignorant agents We define ϵ-ignorant agents as agents whose belief ρ has error (absolute or relative depending on the context) ϵ with respect to the correct belief ρ^*.

For belief, we define both relative and absolute error. This is in contrast with utility, for which this does not make sense in our setting. This is because when one speaks of relative utility, one usually compares it to some default action of "doing nothing" which we do not have.



Definition 5. *We say that a belief ρ has absolute error ϵ with respect to the correct belief ρ^* if for any $t \in \mathbb{N}$, history $\ae_{<t}$ and action a,*

$$\|\rho(\ae_{<t}a) - \rho^*(\ae_{<t}a)\|_{TV} \leq \epsilon \qquad (\star)$$

where $\| \bullet \|_{TV}$ is the total variational distance.

Recall that (on discrete measure spaces where all subsets are measurable) for two distributions (formally two probability measures) μ, ν on $\mathcal{E}$, the total variational distance is defined as

$$\|\mu - \nu\|_{TV} = \sup_{E \subseteq \mathcal{E}} |\mu(E) - \nu(E)|$$

Definition 6. *We say a belief ρ has relative error ϵ with respect to the correct belief ρ^* if for any $t \in \mathbb{N}$, any history $\ae_{<t}$, action a, and percept e,*

$$\frac{1}{1+\epsilon} \leq \frac{\rho(e|\ae_{<t}a)}{\rho^*(e|\ae_{<t}a)} \leq 1 + \epsilon$$

Impatient agents We define impatient agents as agents whose discount factor γ is smaller than the correct discount factor γ^*. This means they have a stronger preference for immediate reward compared to the same reward in the future.

5 Exposition of the results

We now formally describe our results, including statements of the theorems. Proofs and description of the techniques used to prove these results are included in the full version of this paper.

5.1 Performance of ϵ-optimizers can deteriorate

In their paper, Everitt et al. (2016) show that, for modification-independent belief and utility function, if we start with a perfect expected utility maximizer and at any time replace the current policy by the initial policy, the expected discounted utility stays the same. Therefore, later policies cannot be worse than the initial policy and no deterioration happens. We show that in the case of ϵ-optimizers, such a replacement can never decrease the expected discounted utility by more than ϵ (Inequality (2)) but can increase it more, meaning that agent's behaviour can deteriorate with time (Inequality (1)). Specifically, it can deteriorate at an exponential rate, until its actions become arbitrarily bad – that is, until the expected future utility lost is the maximum achievable utility (which is $\frac{1}{1-\gamma}$).

Theorem 7. *Let ρ and u be modification-independent. Consider a self-modifying agent which is an ϵ-optimizer for the empty history. Then, for every $t \geq 1$,*

$$E_{\ae_{<t}}[Q(\ae_{<t}\pi_t(\ae_{<t}))] \geq E_{\ae_{<t}}[Q(\ae_{<t}\pi_1(\ae_{<t}))] - \min(\frac{\epsilon}{\gamma^{t-1}}, \frac{1}{1-\gamma}) \qquad (1)$$

where the expectation is with respect to $\ae_{<t}$ such that the perceptions are distributed according to the belief and the actions are given by $a_i = \pi_i(\ae_{<i})$.

Moreover, for all histories $\ae_{<t}$ given by $a_i = \pi_i(\ae_{<i})$ for which the agent is an ϵ-optimizer it holds that

$$Q(\ae_{<t}\pi_1(\ae_{<t})) + \epsilon \geq Q(\ae_{<t}\pi_t(\ae_{<t})) \qquad (2)$$

Equality in Inequality (1) *can be achieved up to a factor of at most γ.*

The expectation in inequality (1) is necessary as can be demonstrated by the following example. Consider an environment in which the first perception is α with probability $\epsilon(1-\gamma)$ and β otherwise. Regardless of the first perception, the utility in the future is always 1 if the action following this perception is a and 0 if b. An ϵ-optimizing agent which performs action a may choose to self-modify to an agent which performs action a after perception α and b after perception β, thus losing $\frac{\gamma}{1-\gamma}$ utility in the case of perception β, regardless of how small ϵ is.

Setting $\epsilon = 0$ allows us to easily recover Theorem 16 from (Everitt et al. 2016), showing that self-modifications do not impact expected discounted utility gained by perfectly rational agents. This proof is also considerably simpler than the one in the original paper.

If we only care about future discounted utility, this deterioration in performance doesn't need to concern us because it only happens at future times when utility is heavily discounted. From the definition of an ϵ-optimizer, the maximum utility lost is indeed only ϵ. On the other hand, if we care about long-term performance of the agent and have only introduced the discount factor for instrumental reasons (as would likely be the case), the possibility of self-modification becomes a serious problem. The discount factor might be introduced because optimizing the long-term future might be computationally intractable.

5.2 Bounded-knowledge agents are ϵ-optimizers

In this section, we discuss perfect utility maximizers with bounded knowledge and show performance guarantees for such agents. In Section 5.3, we combine these results, show how to relax the assumption of perfect optimization and, most importantly, show how the performance of a bounded-rational agent differs between the cases with and without self-modification.

In Lemma 8, we show that if the agent's estimate of the expected discounted utility is at most ϵ away from the true value, the agent will be a 2ϵ-optimizer. In sections 5.2 to 5.2, we show bounds on how inaccurate the agent's estimate of expected discounted utility can be, thus proving bounds on optimization. In Section 5.2, we proceed differently: we formulate the worst case as a solution of an optimization problem which we then solve analytically.

Lemma 8. *Let A be a (possibly self-modifying) perfect expected utility maximizer with knowledge $\kappa = (u, \rho, \gamma)$. Let $\kappa^* = (u^*, \rho^*, \gamma^*)$ be the correct knowledge. Assume that*

$$|V^\pi_\kappa(\text{\ae}_{<t}) - V^\pi_{\kappa^*}(\text{\ae}_{<t})| \le \epsilon$$

for all policies π and histories $\text{\ae}_{<t}$. Then, A is a 2ϵ-optimizer with respect to κ^*.

ϵ-misaligned agents are ϵ'-optimizers We now consider agents with an inaccurate utility function and derive bounds on ϵ' such that the misaligned agent is an ϵ'-optimizer.

Theorem 9. *Let A be a perfect utility maximizer with utility function u and error ϵ with respect to the correct utility function u^*. Then it is a $\frac{2\epsilon}{1-\gamma}$-optimizer with respect to u^*. Moreover, this bound is tight.*

In the random-error case when for any $\text{\ae}_{<t}$, we randomly choose $u(\text{\ae}_{<t})$ from the set $\{\max(0, u^*(\text{\ae}_{<t}) - \epsilon), \min(1, u^*(\text{\ae}_{<t}) + \epsilon)\}$, we give a simple lower bound that is only a factor 4 away from the upper bound. Consider an environment with only one perception 1 and actions $\{0, 1\}$ and $u^*(1|\text{\ae}_{<t}0) = 1 - 2\epsilon$ and $u^*(1|\text{\ae}_{<t}1) = 1$. With probability $1/4$, it holds that $u(1|\text{\ae}_{<t}1) = u(1|\text{\ae}_{<t}0)$, in which case the agent may take the suboptimal action 0, thus losing 2ϵ in instantaneous utility. At every step, it therefore loses $\epsilon/2$ instantaneous expected utility. In total, it then loses in expectation $\frac{\epsilon}{2(1-\gamma)}$. We have thus proved the following:

Theorem 10. *Let A be a perfect utility maximizer whose utility function u is such that for any $e_t, \text{\ae}_{<t}a_t$, the value $u(e_t|\text{\ae}_{<t}a_t)$ is chosen independently and uniformly from the set $\{\max(0, u^*(e_t|\text{\ae}_{<t}a_t) - \epsilon), \min(1, u^*(e_t|\text{\ae}_{<t}a_t) + \epsilon)\}$. Then, the amount of utility lost is in expectation*

$$\frac{\epsilon}{2(1-\gamma)}$$

ϵ-ignorant agents are ϵ'-optimizers In this section, we discuss agents with inaccurate belief. Theorem 11 gives bounds on the utility lost as a result of the agent having an inaccurate belief. We give an upper bound in terms of the (weaker) absolute error and lower bounds in terms of both absolute and relative error, showing that the upper bound is tight up to factors of 2 and 4 for absolute and relative error.

Theorem 11. *Let A be a perfect expected utility maximizer whose belief ρ has absolute error ϵ with respect to the correct belief ρ^*. Then it is a $(\frac{2}{1-\gamma} - \frac{2}{1-\gamma(1-\epsilon)})$-optimizer with respect to ρ^* and this bound is tight up to a factor of 2. Moreover, if ϵ is the relative error, this bound is tight up to a factor of 4.*

So far, we have considered the worst-case scenario. In the next theorem, we show that in the case of both absolute and relative error, the upper bound is tight up to a constant factor even in the case when the error at each timestep is randomly chosen from the set $\{-\epsilon, \epsilon\}$ (that is, $\rho(e_t|\text{\ae}_{<t}a_t)$ is chosen uniformly from the set $\{\max(0, \rho^*(e_t|\text{\ae}_{<t}a_t) - \epsilon), \min(1, \rho^*(e_t|\text{\ae}_{<t}a_t) + \epsilon)\}$ in the case of absolute error and $\{\frac{\rho^*(e_t|\text{\ae}_{<t}a_t)}{1+\epsilon}, \min(1, (1+\epsilon)\rho^*(e_t|\text{\ae}_{<t}a_t))\}$ in the case of relative error), independently of other timesteps.

Theorem 12. *Let A be a perfect expected utility maximizer whose belief ρ is such that for any $e_t, \text{\ae}_{<t}a_t$, the value $\rho(e_t|\text{\ae}_{<t}a_t)$ is independently for any argument chosen uniformly from the set $\{\max(0, \rho^*(e_t|\text{\ae}_{<t}a_t) - \epsilon), \min(1, \rho^*(e_t|\text{\ae}_{<t}a_t) + \epsilon)\}$ in the case of absolute error and $\{\frac{\rho^*(e_t|\text{\ae}_{<t}a_t)}{1+\epsilon}, \min(1, (1+\epsilon)\rho^*(e_t|\text{\ae}_{<t}a_t))\}$ in the case of relative error.*

Then, in expectation, the amount of expected discounted utility lost is respectively

$$\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\epsilon/8)}$$

$$\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\epsilon/16)}$$

and this is equal to the upper bound for non-random error up to a factor of 16 and 32, respectively.

Impatient agents are ϵ-optimizers In this section, we discuss the case when an agent has an incorrect discount factor and give a bound on the performance of this agent with respect to the correct discount factor. We only consider the case when the agent discounts faster than the correct discount rate – we deem this to be the interesting case as, generally speaking, while optimizing in the long-term might be desirable, it is difficult to achieve, so the agent is likely to optimize in shorter term than desired. Bounds for the other case can be derived by the same method. To simplify the bounds, we define $k = \lceil -\frac{1}{\lg \gamma} \rceil$.

Theorem 13. *Let π_γ and π_{γ^*} be perfect expected utility maximizers with respect to discount factors γ and γ^* respectively for some $\gamma \le \gamma^*$, either with or without the ability to self-modify. Let u,ρ be their utility function and belief. Then*

$$\begin{aligned}|V^{\pi_{\gamma^*}}(\text{\ae}_{<t}) - V^{\pi_\gamma}(\text{\ae}_{<t})| \le &\frac{{\gamma^*}^k + {\gamma^*}^{k-1} - 1}{1-\gamma^*} \\ &- {\gamma^*}^{k-1}\frac{\gamma^k + \gamma^{k-1} - 1}{\gamma^{k-1}(1-\gamma)}\end{aligned}$$

For $\gamma \to 1$, it holds that $\lceil -\frac{1}{\lg\gamma} \rceil \sim -\frac{1}{\lg\gamma}$. This enables us to simplify the previous result to get a good approximation for when γ is close to 1:

$$\frac{{\gamma^*}^k + {\gamma^*}^{k-1} - 1}{1-\gamma^*} - {\gamma^*}^{k-1}\frac{\gamma^k + \gamma^{k-1} - 1}{\gamma^{k-1}(1-\gamma)} \approx \frac{2{\gamma^*}^{\frac{1}{\lg\gamma}} - 1}{1-\gamma^*}$$

5.3 Combining the results

In this section we combine the results from sections 5.1 and 5.2 and present a bound on the utility lost by an agent which is misaligned, ignorant, impatient and has bounded optimization, all at the same time. It is an interesting feature of this bound that the worst-case performance guarantee can in some cases be improved by adjusting its discount rate.

Recall that the functions $f_\bullet$ in the following theorem have been defined in Section 1.1.

Theorem 14. *Let A be an ϵ_o-optimizer for the empty history with either (1) the ability to self-modify and modification-independent utility function and belief, or (2) without the ability to self-modify and with a possibly modification-dependent utility function and belief. Let γ be the agent's discount rate, ϵ_u the error in its utility function wrt. the correct utility function u^* and ϵ_ρ its absolute error in belief with respect to the correct belief ρ^*. Then at timestep t:*

(1) If we let ϵ' be the smallest possible number such that A at time t is an ϵ'-optimizer, then $E_{æ_{<t}}[\epsilon'] \leq f_{opt}(\epsilon_o, \gamma) + f_{util}(\epsilon_u, \gamma) + f_{bel}(\epsilon_\rho, \gamma) + f_{disc}(\gamma, \gamma^*)$ where the expectation is over histories where perceptions are distributed according to ρ^* and actions are given by the agent's policy. Moreover, if $\epsilon_o = 0$, then $\epsilon' \leq f_{opt}(\epsilon_o, \gamma) + f_{util}(\epsilon_u, \gamma) + f_{bel}(\epsilon_\rho, \gamma) + f_{disc}(\gamma, \gamma^*)$ almost certainly.

(2) A will be an ϵ'-optimizer, with respect to the correct discount rate γ^*, where $\epsilon' \leq \epsilon_o + f_{util}(\epsilon_u, \gamma) + f_{bel}(\epsilon_\rho, \gamma) + f_{disc}(\gamma, \gamma^*)$

Moreover, when $\gamma \geq 1/2$, there exists an agent which achieves equality up to a factor of at most 8 and up to a factor of 16 if ϵ_ρ is the relative error.

6 Future work

We propose several directions for future research. In general, it would be interesting to explore the central problem of self-modification safety under different agent and environment models and with different assumptions.

Bounded rationality models. We analyzed a model of bounded-rationality with a strict upper bound on the size of errors (of several kinds). While this shows that even agents guaranteed to have small errors may self-modify in detrimental ways, the analysis would be significantly different for fully stochastic bounded rationality models (e.g. negligible expected errors with non-negligible variance). One such model of interest is Information-Theoretic Bounded Rationality of Ortega et al. (2015).

Awareness of own bounded-rationality. Whatever underlying decision procedure the agent uses somehow *implicitly* takes its ϵ-optimality into account – in particular since the assumed ϵ-optimality depends on the behavior of future agent versions. In our formulation, we do not assume the agent to have *explicit* knowledge of its bounded rationality model and ϵ, which would at least intuitively seem useful to know.

Note, however, that in our framing such explicit knowledge would not be necessarily useful, as any deliberation about it is subject to the same error within ϵ-optimality. Therefore it may be interesting to explore bounded rationality models where the information about own bounded rationality could be explicitly reasoned about (with more precision than e.g. modelling the trajectory of the full environment). Would the agent then be more reluctant to self-modify?

Time horizons and discounting. Avoiding temporal discounting would likely yield results with stronger implications (Section 2). We propose analysing the finite-time undiscounted case, as well as exploring other means of future value aggregation (finite or infinite, as explored by Bostrom (2011)).

Model of self-modification. As noted above, embedded agents with a strong influence on the environment may self-modify by exploiting the environment. However, the extent of this self-modification, and the strength and stability of mechanisms against self-modification (e.g. via modification-dependent utility function) require further research.

Probabilistic analysis. Build stochastic models of agent rationality and self-modification, and perform full probabilistic analysis. This may e.g. inform us about required safety margins. In particular, approaches based on statistical physics and information theory seems to be promising here and have already proven fruitful in analyzing existing optimization problems and algorithms.

7 Acknowledgements

This work was carried out as part of the AI Safety Research Program 2019. We are grateful to Vojtěch Kovařík, Jan Kulveit, and Gavin Leech (in alphabetical order) for their valuable inputs.

References

Armstrong, S.; Sandberg, A.; and Bostrom, N. 2012. Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines* 22. doi:10.1007/s11023-012-9282-2.

Bostrom, N. 2011. Infinite Ethics. *Analysis and Metaphysics* 10: 9–59. URL www.nickbostrom.com/ethics/infinite.pdf.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. USA: Oxford University Press, Inc., 1st edition. ISBN 0199678111.

Demski, A.; and Garrabrant, S. 2019. Embedded Agency. *CoRR* abs/1902.09469. URL http://arxiv.org/abs/1902.09469.

Drexler, K. E. 2019. Reframing superintelligence: Comprehensive AI services as general intelligence. *Future of Humanity Institute, University of Oxford* .

Everitt, T.; Filan, D.; Daswani, M.; and Hutter, M. 2016. Self-Modification of Policy and Utility Function in Rational Agents. *CoRR* abs/1605.03142. URL http://arxiv.org/abs/1605.03142.

Lattimore, T.; and Hutter, M. 2014. General time consistent discounting. *Theoretical Computer Science* 519: 140 – 154. ISSN 0304-3975. doi:https://doi.org/10.1016/j.tcs.2013.09.022. URL http://www.sciencedirect.com/science/article/pii/S0304397513007135. Algorithmic Learning Theory.

Miller, J. D.; Yampolskiy, R.; and Häggström, O. 2020. An AGI Modifying Its Utility Function in Violation of the Orthogonality Thesis. *arXiv preprint arXiv:2003.00812* .

Omohundro, S. M. 2008. The basic AI drives. In *AGI*, volume 171, 483–492.

Ortega, P. A.; Braun, D. A.; Dyer, J.; Kim, K.-E.; and Tishby, N. 2015. Information-Theoretic Bounded Rationality.

Russell, S. 2019. *Human compatible: artificial intelligence and the problem of control*. New York, New York: Viking. ISBN 9780525558613.

Strotz, R. H. 1955. Myopia and inconsistency in dynamic utility maximization. *The review of economic studies* 23(3): 165–180.

Yampolskiy, R. 2012. Leakproofing the Singularity Artificial Intelligence Confinement Problem. *Journal of Consciousness Studies* 19: 194–214.