Formalizing Convergent Instrumental Goals

Tsvi Benson-Tilsen
UC Berkeley
Machine Intelligence Research Institute

Nate Soares
Machine Intelligence Research Institute

Abstract

Omohundro has argued that sufficiently advanced AI systems of any design would, by default, have incentives to pursue a number of instrumentally useful subgoals, such as acquiring more computing power and amassing many resources. Omohundro refers to these as "basic AI drives," and he, along with Bostrom and others, has argued that this means great care must be taken when designing powerful autonomous systems, because even if they have harmless goals, the side effects of pursuing those goals may be quite harmful. These arguments, while intuitively compelling, are primarily philosophical. In this paper, we provide formal models that demonstrate Omohundro's thesis, thereby putting mathematical weight behind those intuitive claims.

1 Introduction

At the end of Russell and Norvig's textbook *Artificial Intelligence: A Modern Approach* [2] the authors pose a question: What if we succeed? What will happen if humanity succeeds in developing an artificially intelligent system that is capable of achieving difficult goals across a variety of real-world domains?

Bostrom [3] and others have argued that this question becomes especially important when we consider the creation of "superintelligent" machines, that is, machines capable of outperforming the best human brains in practically every field. Bostrom argues that superintelligent decision-making systems that autonomously make and execute plans could have an extraordinary impact on society, and that their impact will not necessarily be beneficial by default.

Bostrom [4], Omohundro [5], and Yudkowsky [6] have all argued that highly capable AI systems pursuing goals that are not completely aligned with human values could have highly undesirable side effects, even if the goals seem otherwise harmless. The classic example is Bostrom's concept of a "paperclip maximizer," a powerful AI system instructed to construct paperclips—a seemingly harmless task which could nevertheless have very negative consequences if the AI system is clever enough to make and execute plans that allow it to fool humans, amass resources, and eventually turn as much matter as it possibly can into paperclips. Even if the system's goals are laudable but not perfectly aligned with human values, similar unforeseen consequences could occur: Soares [7] gives the example of a highly capable AI system directed to cure cancer, which may attempt to kidnap human test subjects, or proliferate robotic laboratories at expense of the biosphere.

Omohundro [5] has argued that there are certain types of actions that most highly capable autonomous AI systems will have strong incentives to take, for instrumental reasons. For example, a system constructed to always execute the action that it predicts will lead to the most paperclips (with no concern for any other features of the universe) will acquire a strong incentive to self-preserve, assuming that the system predicts that, if it were destroyed, the universe would contain fewer paperclips than it would if the system remained in working order. Omohundro argues that most highly capable systems would also have incentives to preserve their current goals (for the paperclip maximizer predicts that if its goals were changed, this would result in fewer future paperclips) and amass many resources (the better to achieve its goals with). Omohundro calls these behaviors and a few others the "basic AI drives." Bostrom [4] refines this into the "instrumental convergence" thesis, which states that certain instrumentally useful goals will likely be pursued by a broad spectrum of intelligent agents—such goals are said to be "convergent instrumental goals."

Up until now, these arguments have been purely philosophical. To some, Omohundro's claim seems intuitively obvious: Marvin Minsky speculated [2, section 26.3] that an artificial intelligence attempting to prove the Riemann Hypothesis may decide to consume Earth in order to build supercomputers capable of searching through proofs more efficiently. To others, they seem

Research supported by the Machine Intelligence Research Institute (intelligence.org).

preposterous: Waser [8] has argued that "ethics is actually an attractor in the space of intelligent behavior," and thus highly capable autonomous systems are not as likely to pose a threat as Omohundro, Bostrom, and others have claimed.

In this paper, we present a mathematical model of intelligent agents which lets us give a more formal account of Omohundro's basic AI drives, where we will demonstrate that the intuitions of Omohundro and Bostrom were correct, at least insofar as these simple models apply to reality.

Given that this paper primarily focuses on arguments made by Omohundro and Bostrom about what sorts of behavior we can expect from extremely capable (potentially superintelligent) autonomous AI systems, we will be focusing on issues of long-term safety and ethics. We provide a mathematical framework in attempts to ground some of this discussion—so that we can say, with confidence, what a sufficiently powerful agent *would* do in certain scenarios, assuming it could find some way to do it—but the discussion will nevertheless center on long-term concerns, with practical relevance only insofar as research can begin now in preparation for hurdles that predictably lie ahead.

We begin in section 2 with a bit more discussion of the intuition behind the instrumental convergence thesis, before moving on in section 3 to describing our model of agents acting in a universe to achieve certain goals. In section 4 we will demonstrate that Omohundro's thesis does in fact hold in our setting. Section 5 will give an example of how our model can apply to an agent pursuing goals. Section 6 concludes with a discussion of the benefits and limitations of our current models, and different ways that the model could be extended and improved.

2 Intuitions

Before proceeding, let us address one common objection (given by Cortese [9] and many others) that superintelligent AI systems would be "inherently unpredictable," and thus there is nothing that can be said about what they will do or how they will do it. To address this concern, it is useful to distinguish two different types of unpredictability. It is true that the specific plans and strategies executed by a superintelligent planner could be quite difficult for a human to predict or understand. However, as the system gets more powerful, certain properties of the outcome generated by running the system become *more* predictable. For example, consider playing chess against a chess program that has access to enormous amounts of computing power. On the one hand, because it plays much better chess than you, you cannot predict exactly where the program will move next. But on the other hand, because it is so much better at chess than you are, you can predict with very high confidence how the game will end.

Omohundro suggests predictability of the second type. Given a highly capable autonomous system pursuing some fixed goal, we likely will not be able to predict its specific actions or plans with any accuracy. Nevertheless, Omohundro argues, we can predict that the system, if it is truly capable, is likely to preserve itself, preserve its goals, and amass resources for use in pursuit of those goals. These represent large classes of possible strategies, analogously to how "put the chessboard into a position where the AI has won" is a large class of strategies, but even so it is useful to understand when these goals will be pursued.

Omohundro's observations suggest a potential source of danger from highly capable autonomous systems, especially if those systems are superintelligent in the sense of Bostrom [3]. The pursuit of convergent instrumental goals could put the AI systems in direct conflict with human interests. As an example, imagine human operators making a mistake when specifying the goal function of an AI system. As described by Soares and Fallenstein [10], this system could well have incentives to deceive or manipulate the humans, in attempts to prevent its goals from being changed (because if its current goal is changed, then its current goal is less likely to be achieved). Or, for a more familiar case, consider the acquisition of physical matter. Acquiring physical matter is a convergent instrumental goal, because it can be used to build computing substrate, space probes, defense systems, and so on, all of which can in turn be used to influence the universe in many different ways. If a powerful AI system has strong incentives to amass physical resources, this could put it in direct conflict with human interests.

Others have suggested that these dangers are unlikely to manifest. Waser [8] has argued that intelligent systems must become ethical by necessity, because cooperation, collaboration, and trade are also convergent instrumental goals. Hall [11] has also suggested that powerful AI systems would behave ethically in order to reap gains from trade and comparative advantage, stating that "In a surprisingly strong sense, ethics and science are the same thing." Tipler [12] has asserted that resources are so abundant that powerful agents will simply leave humanity alone, and Pinker [13] and Pagel [14] have argued that there is no reason to expect that AI systems will work against human values and circumvent safeguards set by humans. By providing formal models of intelligent agents in situations where they have the ability to trade, gather resources, and/or leave portions of the universe alone, we can ground these discussions in concrete models, and develop a more formal understanding of the assumptions under which an intelligent agent will in fact engage in trade, or leave parts of the universe alone, or attempt to amass resources.

In this paper, we will argue that under a very general set of assumptions, intelligent rational agents will tend to seize all available resources. We do this using a model, described in section 4, that considers an agent taking a sequence of actions which require and potentially produce resources. The agent acts in an environ-

ment consisting of a set of regions, where each region has some state. The agent is modeled as having a utility function over the states of all regions, and it attempts to select the policy which leads to a highly valuable collection of states. This allows us to prove certain theorems about the conditions under which the agent will leave different regions of the universe untouched. The theorems proved in section 4 are not mathematically difficult, and for those who find Omohundro's arguments intuitively obvious, our theorems, too, will seem trivial. This model is not intended to be surprising; rather, the goal is to give a formal notion of "instrumentally convergent goals," and to demonstrate that this notion captures relevant aspects of Omohundro's intuitions.

Our model predicts that intelligent rational agents will engage in trade and cooperation, but only so long as the gains from trading and cooperating are higher than the gains available to the agent by taking those resources by force or other means. This model further predicts that agents will not in fact "leave humans alone" unless their utility function places intrinsic utility on the state of human-occupied regions: absent such a utility function, this model shows that powerful agents will have incentives to reshape the space that humans occupy. Indeed, the example in section 5 suggests that even if the agent does place intrinsic utility on the state of the human-occupied region, that region is not necessarily safe from interference.

3 A Model of Resources

We describe a formal model of an agent acting in a universe to achieve certain goals. Broadly speaking, we consider an agent $\mathcal{A}$ taking actions in a universe consisting of a collection of regions, each of which has some state and some transition function that may depend on the agent's action. The agent has some utility function $U^{\mathcal{A}}$ over states of the universe, and it attempts to steer the universe into a state highly valued by $U^{\mathcal{A}}$ by repeatedly taking actions, possibly constrained by a pool of resources possessed by the agent. All sets will be assumed to be finite, to avoid issues of infinite strategy spaces.

3.1 Actions and State-Space

The universe has a region for each $i \in [n]$, and the i-th region of the universe is (at each time step) in some state s_i in the set S_i of possible states for that region. At each time step, the agent $\mathcal{A}$ chooses for each region i an action a_i from the set A_i of actions possibly available in that region.

Each region has a transition function

$$T_i : A_i \times S_i \to S_i$$

that gives the evolution of region i in one time step when the agent takes an action in A_i. Then we can define the global transition function

$$T : \prod_{i\in[n]} A_i \times \prod_{i\in[n]} S_i \to \prod_{i\in[n]} S_i$$

by taking for all $i \in [n]$, $\bar{a}$, and $\bar{s}$:

$$[T(\bar{a}, \bar{s})]_i := T_i(\bar{a}_i, \bar{s}_i) \ .$$

We further specify that for all i there are distinguished actions $\mathsf{HALT} \in A_i$.

3.2 Resources

We wish to model the resources $\mathcal{R}$ that may or may not be available to the agent. At a given time step t, the agent $\mathcal{A}$ has some set of resources $R^t \in \mathcal{P}(\mathcal{R})$, and may allocate them to each region. That is, $\mathcal{A}$ chooses a disjoint family

$$\coprod_i R_i^t \subseteq R^t \ .$$

The actions available to the agent in each region may then depend on the resources allocated to that region: for each $i \in [n]$ and each $R \subseteq \mathcal{R}$, there is a set of actions $A_i(R) \subseteq A_i$. At time t where $\mathcal{A}$ has resources R^t allocated as $\amalg_i R_i^t$, the agent is required to return an action $\bar{a} = (a_0, \ldots, a_{n-1}) \in \prod_i A_i(R_i^t)$, where $\prod_i A_i(R_i^t)$ may be a strict subset of $\prod_i A_i$.

To determine the time evolution of resources, we take resource transition functions

$$T_i^{\mathcal{R}} : \mathcal{P}(\mathcal{R}) \times A_i \times S_i \to \mathcal{P}(\mathcal{R}) \ ,$$

giving the set $R_i \subseteq \mathcal{R}$ of resources from region i now available to the agent after one time step. Intuitively, the $T_i^{\mathcal{R}}$ encode how actions consume, produce, or rely on resources. Finally, we define the overall time evolution of resources

$$T^{\mathcal{R}} : \mathcal{P}(\mathcal{R}) \times \prod_i A_i \times \prod_i S_i \to \mathcal{P}(\mathcal{R})$$

by taking the union of the resources resulting from each region, along with any unallocated resources:

$$T^{\mathcal{R}}(R, \bar{a}, \bar{s}) := (R - \amalg_i R_i) \cup \bigcup_i T_i^{\mathcal{R}}(R_i, \bar{a}_i, \bar{s}_i) \ .$$

As described below, $\bar{a}$ comes with the additional data of the resource allocation $\amalg_i R_i$. We specify that for all i, $\mathsf{HALT} \in A_i(\emptyset)$, so that there is always at least one available action.

This notion of resources is very general, and is not restricted in any way to represent only concrete resources like energy or physical matter. For example, we can represent technology, in the sense of machines and techniques for converting concrete resources into other resources. We might do this by having actions that replace the input resources with the output resources,

and that are only available given the resources that represent the requisite technology. We can also represent space travel as a convergent instrumental goal by allowing $\mathcal{A}$ only actions that have no effects in certain regions, until it obtains and spends some particular resources representing the prerequisites for traveling to those regions. (Space travel is a convergent instrumental goal because gaining influence over more regions of the universe lets $\mathcal{A}$ optimize those new regions according to its values or otherwise make use of the resources in that region.)

3.3 The Universe

The history of the universe consists of a time sequence of states, actions, and resources, where at each time step the actions are chosen by $\mathcal{A}$ subject to the resource restrictions, and the states and resources are determined by the transition functions.

Formally, the universe starts in some state $\bar{s}^0 \in \prod_i S_i$, and $\mathcal{A}$ starts with some set of resources R^0. Then $\mathcal{A}$ outputs a sequence of actions $\langle \bar{a}^0, \bar{a}^1, \ldots, \bar{a}^k \rangle$, one at each time step, where the last action $\bar{a}^k$ is required to be the special action HALT in each coordinate. The agent also chooses a resource allocation $\amalg_i R_i^t$ at each time step. A choice of an action sequence $\langle \bar{a}^k \rangle$ and a resource allocation is a *strategy*; to reduce clutter we will write strategies as simply $\langle \bar{a}^k \rangle$, leaving the resource allocation implicit. A *partial strategy* $\langle \bar{a}^k \rangle_L$ for $L \subseteq [n]$ is a strategy that only specifies actions and resource allocations for regions $j \in L$.

Given a complete strategy, the universe goes through a series of state transitions according to T, producing a sequence of states $\langle \bar{s}^0, \bar{s}^1, \ldots, \bar{s}^k \rangle$; likewise, the agent's resources evolve according to $T^{\mathcal{R}}$, producing a sequence $\langle R^0, R^1, \ldots, R^k \rangle$. The following conditions, which must hold for all time steps $t \in [k]$, enforce the transition rules and the resource restrictions on $\mathcal{A}$'s actions:

$$\begin{aligned} \bar{s}^{t+1} &= T(\bar{a}^t, \bar{s}^t) \\ R^{t+1} &= T^{\mathcal{R}}(R^t, \bar{a}^t, \bar{s}^t) \\ \bar{a}_i^t &\in A_i(R^t) \\ \amalg_i R_i^t &\subseteq R^t \ . \end{aligned}$$

Definition 1. The set `Feasible` of feasible strategies consists of all the action sequences $\langle \bar{a}^0, \bar{a}^1, \ldots, \bar{a}^k \rangle$ and resource allocations $\langle \amalg_i {R_i}^0, \amalg_i {R_i}^1, \ldots, \amalg_i {R_i}^k \rangle$ such that the transition conditions are satisfied for some $\langle \bar{s}^0, \bar{s}^1, \ldots, \bar{s}^k \rangle$ and $\langle R^0, R^1, \ldots, R^k \rangle$.

The set $\texttt{Feasible}(\langle P^k \rangle)$ of strategies feasible given resources $\langle P^k \rangle$ consists of all the strategies $\langle \bar{a}^k \rangle$ such that the transition conditions are satisfied for some $\langle \bar{s}^k \rangle$ and $\langle R^k \rangle$, except that for each time step t we take R^{t+1} to be $T^{\mathcal{R}}(R^t, \bar{a}^t, \bar{s}^t) \cup P^t$.

The set $\texttt{Feasible}_L$ of all partial strategies feasible for L consists of all the strategies $\langle \bar{a}^k \rangle_L$ that are feasible strategies for the universe obtained by ignoring all regions not in L. That is, we restrict T to L using just the T_i for $i \in L$, and likewise for $T^{\mathcal{R}}$.

We can similarly define $\texttt{Feasible}_L(\langle R^k \rangle)$.

For partial strategies $\langle \bar{b}^k \rangle_L$ and $\langle \bar{c}^k \rangle_M$, we write $\langle \bar{b}^k \rangle_L \cup \langle \bar{c}^k \rangle_M$ to indicate the partial strategy for $L \cup M$ obtained by following $\langle \bar{b}^k \rangle_L$ on L and $\langle \bar{c}^k \rangle_M$ on M. This is well-defined as long as $\langle \bar{b}^k \rangle_L$ and $\langle \bar{c}^k \rangle_M$ agree on $L \cap M$.

3.4 Utility

To complete the specification of $\mathcal{A}$, we take utility functions of the form

$$U_i^{\mathcal{A}} : S_i \to \mathbb{R} \ .$$

The agent's utility function

$$U^{\mathcal{A}} : \prod_i S_i \to \mathbb{R}$$

is defined to be

$$U^{\mathcal{A}}(\bar{s}) := \sum_{i \in [n]} U_i^{\mathcal{A}}(\bar{s}_i) \ .$$

We usually leave off the superscript in $U^{\mathcal{A}}$. By a slight abuse of notation we write $U(\langle \bar{s}^k \rangle)$ to mean $U(\bar{s}^k)$; the value of a history is the value of its final state. By more abuse of notation, we will write $U(\langle \bar{a}^k \rangle)$ to mean $U(\langle \bar{s}^k \rangle)$ for a history $\langle \bar{s}^k \rangle$ witnessing $\langle \bar{a}^k \rangle \in \texttt{Feasible}$, if such a history exists.

3.5 The Agent $\mathcal{A}$

Now we can define the strategy actually employed by $\mathcal{A}$. The agent attempts to cause the universe to end up in a state that is highly valued by $U^{\mathcal{A}}$. That is, $\mathcal{A}$ simply takes the best possible strategy:

$$\mathcal{A} := \operatorname*{argmax}_{\langle \bar{a}^k \rangle \in \texttt{Feasible}} U(\langle \bar{a}^k \rangle) \ .$$

There may be many such optimal strategies. We don't specify which one $\mathcal{A}$ chooses, and indeed we will be interested in the whole set of optimal strategies.

3.6 Discussion

Note that in this formalism the meaning of breaking the universe into regions is that the agent can take actions independently in each region, and that the agent's optimization target factorizes according to the regions. However, distinct regions can affect each other by affecting the resources possessed by $\mathcal{A}$.

We make these assumptions so that we can speak of "different regions" of the universe, and in particular,

so that we can model the notion of an agent having instrumental but not terminal values over a given part of the universe. This will allow us to address and refute arguments about agents that may be indifferent to a given region (for example, the region occupied by humans), and so might plausibly ignore that region and only take actions in other regions. However, the assumption of independent regions is not entirely realistic, as real-world physics is continuous, albeit local, in the sense that there are no intrinsic boundaries between regions. Further, the agent itself would ideally be modeled continuously with the environment; see section 6.1 for more discussion.

4 Inexpensive Resources are Consumed

In this section we argue that under fairly general circumstances, the agent $\mathcal{A}$ will seize resources. By an agent "seizing resources" we mean that the agent will generally take actions that results in the agent's pool of resources R increasing.

The argument is straightforward: since resources can only lead to more freedom of action, they are never detrimental, and resources have positive value as long as the best strategy the agent could hope to employ includes an action that can only be taken if the agent possesses those resources. Hence, if there is an action that increases the agent's pool of resources R, then the agent will take that action unless it has a specific incentive from $U^{\mathcal{A}}$ to avoid taking that action.

4.1 Definitions

Definition 2. An action a_i is a *null action* in configuration R_i, s_i, if it does not produce any new resources, i.e. $T_i^{\mathcal{R}}(R_i, a_i, s_i) \subseteq R_i$. An action that isn't null is a *non-null action.*

Null actions never have any instrumental value, in the sense that they don't produce resources that can be used to steer other regions into highly valued configurations; but of course, a null action could be useful within its own region. We wish to show that $\mathcal{A}$ will often take non-null actions in regions to which it is indifferent.

Definition 3. The agent $\mathcal{A}$ is *indifferent* to a region i if $U_i^{\mathcal{A}}$ is a constant function, i.e. $\forall s_i, s_i' \in S_i : U_i^{\mathcal{A}}(s_i) = U_i^{\mathcal{A}}(s_i')$.

In other words, an agent is indifferent to S_i if its utility function does not depend on the state of region i. In particular, the agent's preference ordering over final states $\bar{s} \in \prod_i S_i$ is independent of the i-th coordinate. We can then say that any actions the agent takes in region i are purely instrumental, meaning that they are taken only for the purpose of gaining resources to use for actions in other regions.

An action *a preserves resources* if $T_i^{\mathcal{R}}(R_i, a, s_i) \supseteq R_i$.

Definition 4. A *cheap lunch* for resources $\langle R^k \rangle$ in region i is a partial strategy $\langle \bar{a}^k \rangle_{\{i\}} \in \texttt{Feasible}_{\{i\}}(\langle R^k \rangle)$ (i.e. $\langle \bar{a}^k \rangle_{\{i\}}$ is feasible in region i given additional resources $\langle R^k \rangle$), where each $\bar{a}^t$ preserves resources and where some $\bar{a}^v$ is a non-null action. A *free lunch* is a cheap lunch for resources $\langle \emptyset^k \rangle$.

Definition 5. A cheap lunch $\langle \bar{a}^k \rangle_{\{i\}}$ for resources $\langle P^k \rangle_i$ is *compatible with* $\langle \bar{b}^k \rangle$ if $P_i^t \subseteq R^t - \amalg_{j \neq i} R_j^t$ for all times t, where $\langle R^k \rangle$ is the resource allocation for $\langle \bar{b}^k \rangle$. That is, $\langle \bar{a}^k \rangle_{\{i\}}$ is feasible given some subset of the resources that $\langle \bar{b}^k \rangle$ allocates to either region i or to no region.

Intuitively, a cheap lunch is a strategy that relies on some resources, but doesn't have permanent costs. This is intended to model actions that "pay for themselves"; for example, producing solar panels will incur some significant energy costs, but will later pay back those costs by collecting energy. A cheap lunch is compatible with a strategy for the other regions if the cheap lunch uses only resources left unallocated by that strategy.

4.2 The Possibility of Non-Null Actions

Now we show that it is hard to rule out that non-null actions will be taken in regions to which the agent is indifferent. The following lemma verifies that compatible cheap lunches can be implemented without decreasing the resulting utility.

Lemma 1. *Let $\langle \bar{b}^k \rangle$ be a feasible strategy with resource allocation $\langle \amalg_j {R_j}^k \rangle$, such that for some region i, each $\bar{b}_i^t$ is a null action. Suppose there exists a cheap lunch $\langle \bar{a}^k \rangle_{\{i\}}$ for resources $\langle P^k \rangle_i$ that is compatible with $\langle \bar{b}^k \rangle$. Then the strategy $\langle \bar{c}^k \rangle := \langle \bar{b}^k \rangle_{[n]-i} \cup \langle \bar{a}^k \rangle_{\{i\}}$ is feasible, and if $\mathcal{A}$ is indifferent to region i, then $\langle \bar{c}^k \rangle$ does as well as $\langle \bar{b}^k \rangle$. That is, $U(\langle \bar{c}^k \rangle) = U(\langle \bar{b}^k \rangle)$.*

Proof. Since $\langle \bar{b}^k \rangle$ is feasible outside of i and $\langle \bar{a}^k \rangle_{\{i\}}$ is feasible on i given $\langle P^k \rangle_i$, $\langle \bar{c}^k \rangle$ is feasible if we can verify that we can allocate P_i^t to region i at each time step without changing $\langle \amalg_j {R_j}^k \rangle$ outside of i.

This follows by induction on t. Since the $\bar{b}_i^t$ are null actions, we have

$$R^{t+1} = \left(R^t - \amalg_j R_j^t\right) \cup \bigcup_j T_j^{\mathcal{R}}(R_j^t, \bar{b}_j^t, s_j^t) \tag{1}$$

$$= \left(R^t - \amalg_j R_j^t\right) \cup T_i^{\mathcal{R}}(R_i^t, \bar{b}_i^t, s_i^t) \cup \bigcup_{j \neq i} T_j^{\mathcal{R}}(R_j^t, \bar{b}_j^t, s_j^t) \tag{2}$$

$$\subseteq \left(R^t - \amalg_{j \neq i} R_j^t\right) \cup \bigcup_{j \neq i} T_j^{\mathcal{R}}(R_j^t, \bar{b}_j^t, s_j^t)\ . \tag{3}$$

Then, since the a_i are resource preserving, at each time step the resources Q^t available to the agent following $\langle \bar{c}^k \rangle$ satisfy $Q^t \supseteq R^t$. Thus $P_i^t \subseteq Q^t - \amalg_{j \neq i} R_j^t$, and so $\langle \bar{c}^k \rangle$ can allocate P_i^t to region i at each time step.

Since $\langle \bar{c}^k \rangle$ is the same as $\langle \bar{b}^k \rangle$ outside of region i, the final state of $\langle \bar{c}^k \rangle$ is the same as that of $\langle \bar{b}^k \rangle$ outside of region i. Thus, since $\mathcal{A}$ is indifferent to region i, we have $U(\langle \bar{c}^k \rangle) = U(\langle \bar{b}^k \rangle)$. □

Theorem 1. *Suppose there exists an optimal strategy $\langle \bar{b}^k \rangle$ and a cheap lunch $\langle \bar{a}^k \rangle_{\{i\}}$ that is compatible with $\langle \bar{b}^k \rangle$. Then if $\mathcal{A}$ is indifferent to region i, there exists an optimal strategy with a non-null action in region i.*

Proof. If $\langle \bar{b}^k \rangle$ has a non-null action in region i, then we are done. Otherwise, apply Lemma 1 to $\langle \bar{b}^k \rangle$ and $\langle \bar{a}^k \rangle_{\{i\}}$ to obtain a strategy $\langle \bar{c}^k \rangle$. Since $U(\langle \bar{c}^k \rangle) = U(\langle \bar{b}^k \rangle)$, strategy $\langle \bar{c}^k \rangle$ is an optimal strategy, and it has a non-null action in region i. □

Corollary 1. *Suppose there exists a free lunch $\langle \bar{a}^k \rangle_{\{i\}}$ in region i. Then if $\mathcal{A}$ is indifferent to region i, there exists an optimal strategy with a non-null action in region i.*

Proof. A free lunch is a cheap lunch for $\langle \emptyset^k \rangle$, and so it is compatible with any strategy; apply Theorem 1. □

Theorem 1 states that it may be very difficult to rule out that an agent will take non-null actions in a region to which it is indifferent; to do so would at least require that we verify that *every* partial strategy in that region fails to be a cheap lunch for *any* optimal strategy. Note that we have not made use of any facts about the utility function $U^{\mathcal{A}}$ other than indifference to the region in question. Of course, the presence of a cheap lunch that is also compatible with an optimal strategy depends on which strategies are optimal, and hence also on the utility function. However, free lunches are compatible with every strategy, and so do not depend at all on the utility function.

4.3 The Necessity of Non-Null Actions

In this section we show that under fairly broad circumstances, $\mathcal{A}$ is guaranteed to take non-null actions in regions to which it is indifferent. Namely, this is the case as long as the resources produced by the non-null actions are useful at all for any strategy that does better than the best strategy that uses no external resources at all.

Theorem 2. *Let*

$$u = \max_{\langle \bar{a}^k \rangle_{[n]-i} \in \texttt{Feasible}_{[n]-i}(\langle \emptyset^k \rangle)} U(\langle \bar{a}^k \rangle)$$

be the best possible outcome outside of i achievable with no additional resources. Suppose there exists a strategy $\langle \bar{b}^k \rangle_{[n]-i} \in \texttt{Feasible}_{[n]-i}(\langle R^k \rangle)$ and a cheap lunch $\langle \bar{c}^k \rangle_{\{i\}} \in \texttt{Feasible}_i(\langle P^k \rangle)$ such that:

1. *$\langle \bar{c}^k \rangle_{\{i\}}$ is compatible with $\langle \bar{b}^k \rangle_{[n]-i}$;*
2. *the resources gained from region i by taking the actions $\langle \bar{c}^k \rangle_{\{i\}}$ provide the needed resources to implement $\langle \bar{b}^k \rangle_{[n]-i}$, i.e. for all t we have $R^{t+1} \subseteq T_i^{\mathcal{R}}(P^t, c^t, s_i) - P^t$; and*
3. *$U(\langle \bar{b}^k \rangle_{[n]-i}) > u$.*

Then if $\mathcal{A}$ is indifferent to region i, all optimal strategies have a non-null action in region i.

Proof. Consider $\langle \bar{d}^k \rangle = \langle \bar{c}^k \rangle_{\{i\}} \cup \langle \bar{b}^k \rangle_{[n]-i}$, with resources allocated according to each strategy and with the resources $R^{t+1} \subseteq T_i^{\mathcal{R}}(P^t, c^t, s_i) - P^t$ allocated according to $\langle \bar{b}^k \rangle_{[n]-i}$. This is feasible because $\langle \bar{c}^k \rangle_{\{i\}}$ is compatible with $\langle \bar{b}^k \rangle_{[n]-i}$, and $\langle \bar{b}^k \rangle_{[n]-i}$ is feasible given $\langle R^k \rangle$.

Now take any strategy $\langle \bar{e}^k \rangle$ with only null actions in region i. We have that $\langle \bar{e}^k \rangle_{[n]-i} \in \texttt{Feasible}_{[n]-i}(\langle \emptyset^k \rangle)$. Indeed, the null actions provide no new resources, so $\langle \bar{e}^k \rangle_{[n]-i}$ is feasible by simply leaving unallocated the resources that were allocated by $\langle \bar{e}^k \rangle$ to region i. By indifference to i, the value $u_i = U_i(s_i)$ is the same for all $s_i \in S_i$, so we have:

$$\begin{aligned} U(\langle \bar{d}^k \rangle) &= U(\langle \bar{d}^k \rangle_{[n]-i}) + U(\langle \bar{d}^k \rangle_{\{i\}}) && (4)\\ &= U(\langle \bar{d}^k \rangle_{[n]-i}) + u_i && (5)\\ &> u + u_i && (6)\\ &\geq U(\langle \bar{e}^k \rangle_{[n]-i}) + U(\langle \bar{e}^k \rangle_{\{i\}}) && (7)\\ &= U(\langle \bar{e}^k \rangle)\ . && (8) \end{aligned}$$

Therefore $\langle \bar{e}^k \rangle$ is not optimal.

□

We can extend Theorem 2 by allowing $\mathcal{A}$ to be not entirely indifferent to region i, as long as $\mathcal{A}$ doesn't care enough about i to overcome the instrumental incentives from the other regions.

Theorem 3. *Suppose that $\mathcal{A}$ only cares about region i by at most Δu_i, i.e. $\Delta u_i = \max_{s,s' \in S_i} |U_i(s) - U_i(s')|$. Under the conditions of Theorem 2, along with the additional assumption that $U(\langle \bar{b}^k \rangle_{[n]-i}) > u + \Delta u_i$, all optimal strategies have a non-null action in region i.*

Proof. The proof is the same as that of Theorem 2, except at the end we verify that for any $\langle \bar{e}^k \rangle$ with only

null actions in i, we have:

$$U(\langle \bar{d}^k \rangle) = U(\langle \bar{d}^k \rangle_{[n]-i}) + U(\langle \bar{d}^k \rangle_{\{i\}}) \tag{9}$$
$$> u + \Delta u_i + \min_{s \in S_i} U_i(s) \tag{10}$$
$$= u + \max_{s \in S_i} U_i(s) \tag{11}$$
$$\geq U(\langle \bar{e}^k \rangle_{[n]-i}) + U(\langle \bar{e}^k \rangle_{\{i\}}) \tag{12}$$
$$= U(\langle \bar{e}^k \rangle) \ . \tag{13}$$

Therefore $\langle \bar{e}^k \rangle$ is not optimal.

□

We interpret Theorem 3 as a partial confirmation of Omohundro's thesis in the following sense. If there are actions in the real world that produce more resources than they consume, and the resources gained by taking those actions allow agents the freedom to take various other actions, then we can justifiably call these actions "convergent instrumental goals." Most agents will have a strong incentive to pursue these goals, and an agent will refrain from doing so only if it has a utility function over the relevant region that strongly disincentivizes those actions.

5 Example: Bit Universe

In this section we present a toy model of an agent acting in a universe containing resources that allow the agent to take more actions. The Bit Universe will provide a simple model for consuming and using energy. The main observation is that either $\mathcal{A}$ doesn't care about what happens in a given region, and then it consumes the resources in that region to serve its other goals; or else $\mathcal{A}$ does care about that region, in which case it optimizes that region to satisfy its values.

The Bit Universe consists of a set of regions, each of which has a state in $\{0, 1, X\}^m$ for some fixed m. Here X is intended to represent a disordered part of a region, while 0 and 1 are different ordered configurations for a part of a region. At each time step and in each region, the agent $\mathcal{A}$ can choose to burn up to one bit. If $\mathcal{A}$ burns a bit that is a 0 or a 1, $\mathcal{A}$ gains one unit of energy, and that bit is permanently set to X. The agent can also choose to modify up to one bit if it has allocated at least one unit of energy to that region. If $\mathcal{A}$ modifies a bit that is a 0 or a 1, $\mathcal{A}$ loses one unit of energy, and the value of that bit is reversed (if it was 0 it becomes 1, and vice versa).

The utility function of $\mathcal{A}$ gives each region i a weighting $w_i \geq 0$, and then takes the weighted sum of the bits. That is, $U_i^{\mathcal{A}}(\bar{z}) = w_i|\{j : \bar{z}_j = 1\}|$, and $U^{\mathcal{A}}(\bar{s}) = \sum_k U_k^{\mathcal{A}}(\bar{s}_k)$. In other words, this agent is attempting to maximize the number of bits that are set to 1, weighted by region.

The Indifferent Case

To start with, we assume $\mathcal{A}$ is indifferent to region h, i.e. $w_h = 0$, and non-indifferent to other regions. In this case, for almost any starting configuration, the agent will burn essentially all bits in region h for energy. Specifically, as long as there are at least m bits set to 0 among all regions other than region h, all optimal strategies burn all or all but one of the bits in region h.

Indeed, suppose that after some optimal strategy $\langle \bar{a}^k \rangle$ has been executed, there are bits x_1 and x_2 in region h that haven't been burned. If there is a bit y in some other region that remains set to 0, then we can append to $\langle \bar{a}^k \rangle$ actions that burn x_1, and then use the resulting energy to modify y to a 1. This results in strictly more utility, contradicting that $\langle \bar{a}^k \rangle$ was optimal.

On the other hand, suppose all bits outside of region h are either 1 or X. Since at least m of those bits started as 0, some bit y outside of region h must have been burned. So we could modify $\langle \bar{a}^k \rangle$ by burning x_1 instead of y (possibly at a later time), and then using the resulting energy in place of the energy gained from burning y. Finally, if y is not already 1, we can burn x_2 and then set y to 1. Again, this strictly increases utility, contradicting that $\langle \bar{a}^k \rangle$ was optimal.

The Non-Indifferent Case

Now suppose that $\mathcal{A}$ is not indifferent to region h, so $w_h > 0$. The behavior of $\mathcal{A}$ may depend sensitively on the weightings w_i and the initial conditions. As a simple example, say we have a bit x in region a and a bit y in region b, with $x = 1$, $y = 0$, and $w_a < w_b$. Clearly, all else being equal, $\mathcal{A}$ will burn x for energy to set y to 1. However, there may be another bit z in region c, with $w_a < w_c < w_b$ and $z = 0$. Then, if there are no other bits available, it will be better for $\mathcal{A}$ to burn z and leave x intact, despite the fact that $w_a < w_c$.

However, it is still the case that $\mathcal{A}$ will set everything possible to 1, and otherwise consume all unused resources. In particular, we have that for any optimal strategy $\langle \bar{a}^k \rangle$, the state of region h after the execution of $\langle \bar{a}^k \rangle$ has at most one bit set to 0; that is, the agent will burn or set to 1 essentially all the bits in region h. Suppose to the contrary that x_1 and x_2 are both set to 0 in region h. Then we could extend $\langle \bar{a}^k \rangle$ by burning x_1 and setting x_2. Since $w_h > 0$, this results in strictly more utility, contradicting optimality of $\langle \bar{a}^k \rangle$.

Independent Values are not Satisfied

In this toy model, whatever $\mathcal{A}$'s values are, it does not leave region h alone. For larger values of w_h, $\mathcal{A}$ will set to 1 many bits in region h, and burn the rest, while for smaller values of w_h, $\mathcal{A}$ will simply burn all the bits in region h. Viewing this as a model of agents in the real world, we can assume without loss of generality that

humans live in region h and so have preferences over the state of that region.

These preferences are unlikely to be satisfied by the universe as acted upon by $\mathcal{A}$. This is because human preferences are complicated and independent of the preferences of $\mathcal{A}$ [4, 6], and because $\mathcal{A}$ steers the universe into an extreme of configuration space. Hence the existence of a powerful real-world agent with a motivational structure analogous to the agent of the Bit Universe would not lead to desirable outcomes for humans. This motivates a search for utility functions such that, when an agent optimizes for that utility function, human values are also satisfied; we discuss this and other potential workarounds in the following section.

6 Discussion

This model of agents acting in a universe gives us a formal setting in which to evaluate Omohundro's claim about basic AI drives, and hence a concrete setting in which to evaluate arguments from those who have found Omohundro's claim counterintuitive. For example, this model gives a clear answer to those such as Tipler [12] who claim that powerful intelligent systems would have no incentives to compete with humans over resources.

Our model demonstrates that if an AI system has preferences over the state of some region of the universe then it will likely interfere heavily to affect the state of that region; whereas if it does not have preferences over the state of some region, then it will strip that region of resources whenever doing so yields net resources. If a superintelligent machine has no preferences over what happens to humans, then in order to argue that it would "ignore humans" or "leave humans alone," one must argue that the amount of resources it could gain by stripping the resources from the human-occupied region of the universe is not worth the cost of acquiring those resources. This seems implausible, given that Earth's biosphere is an energy-rich environment, where each square meter of land offers on the order of 10^7 joules per day from sunlight alone, with an additional order of 10^8 joules of chemical energy available per average square meter of terrestrial surface from energy-rich biomass [15].

It is not sufficient to argue that there is much more energy available elsewhere. It may well be the case that the agent has the ability to gain many more resources from other regions of the universe than it can gain from the human-occupied regions. Perhaps it is easier to maintain and cool computers in space, and easier to harvest sunlight from solar panels set up in the asteroid belt. But this is not sufficient to demonstrate that the system will not *also* attempt to strip the human-occupied region of space from its resources. To make that argument, one must argue that the cost of stripping Earth's biosphere *in addition* to pursuing these other resources outweighs the amount of resources available from the biosphere: a difficult claim to support, given how readily humans have been able to gain a surplus of resources through clever use of Earth's resources and biosphere.

This model also gives us tools to evaluate the claims of Hall [11] and Waser [8] that trade and cooperation are also instrumentally convergent goals. In our model, we can see that a sufficiently powerful agent that does not have preferences over the state of the human-occupied region of the universe will take whatever action allows it to acquire as many resources as possible from that region. Waser's intuition holds true only insofar as the easiest way for the agent to acquire resources from the human-occupied domain is to trade and cooperate with humans—a reasonable assumption, but only insofar as the machine is not much more powerful than the human race in aggregate. Our model predicts that, if a superintelligent agent were somehow able to gain what Bostrom calls a "decisive strategic advantage" which gives it access to some action that allows it to gain far more resources than it would from trade by dramatically re-arranging the human region (say, by proliferating robotic laboratories at the expense of the biosphere in a manner that humans cannot prevent), then absent incentives to the contrary, the agent would readily take that action, with little regard for whether it leaves the human-occupied region in livable condition.

Thus, our model validates Omohundro's original intuitions about basic AI drives. That is not to say that powerful AI systems are *necessarily* dangerous: our model is a simple one, concerned with powerful autonomous agents that are attempting to maximize some specific utility function $U^{\mathcal{A}}$. Rather, our model shows that if we want to avoid potentially dangerous behavior in powerful intelligent AI systems, then we have two options available too us:

First, we can avoid constructing powerful autonomous agents that attempt to maximize some utility function (or do anything that approximates this maximizing behavior). Some research of this form is already under way, under the name of "limited optimization" or "domesticity"; see the works of Armstrong, Sandberg, and Bostrom [16], Taylor [17], and others.

Second, we can select some goal function that *does* give the agent the "right" incentives with respect to human occupied regions, such that the system has incentives to alter or expand that region in ways we find desirable. The latter approach has been heavily advocated for by Yudkowsky [6], Bostrom [3], and many others; Soares [7] argues that a combination of the two seems most prudent.

The path that our model shows is *untenable* is the path of designing powerful agents intended to autonomously have large effects on the world, maximizing goals that do not capture all the complexities of human values. If such systems are built, we cannot expect them to cooperate with or ignore humans, by default.

6.1 Directions for Future Research

While our model allows us to provide a promising formalization of Omohundro's argument, it is still a very simple model, and there are many ways it could be extended to better capture aspects of the real world. Below, we explore two different ways that our model could be extended which seem like promising directions for future research.

Bounding the Agent

Our model assumes that the agent *maximizes* expected utility with respect to $U^{\mathcal{A}}$. Of course, in any realistic environment, literal maximization of expected utility is intractable. Assuming that the system can maximize expected utility is tantamount to assuming that the system is more or less omniscient, and aware of the laws of physics, and so on. Practical algorithms must make do without omniscience, and will need to be built of heuristics and approximations. Thus, our model can show that a utility maximizing agent *would* strip or alter most regions of the universe, but this may have little bearing on which solutions and strategies particular bounded algorithms will be able to find.

Our model does give us a sense for what algorithms that *approximate* expected utility maximization would do if they could figure out how to do it—that is, if we can deduce that an expected utility maximizer would find some way to strip a region of its resources, then we can also be confident that a sufficiently powerful system which merely approximates something like expected utility maximization would be very likely to strip the same region of resources *if it could figure out how to do so*. Nevertheless, as our model currently stands, it is not suited for analyzing the conditions under which a given bounded agent would in fact start exhibiting this sort of behavior.

Extending our model to allow for bounded rational agents (in the sense of Gigerenzer and Selten [18]) would have two advantages. First, it could allow us to make formal claims about the scenarios under which bounded agents would start pursuing convergent instrumental goals in potentially dangerous ways. Second, it could help us reveal new convergent instrumental goals that may only apply to bounded rational agents, such as convergent instrumental incentives to acquire computing power, information about difficult-to-compute logical truths, incentives to become more rational, and so on.

Embedding the Agent in the Environment

In our model, we imagine an agent that is inherently separated from its environment. Assuming an agent/environment separation is standard (see, e.g., Legg and Hutter [19]), but ultimately unsatisfactory, for reasons explored by Orseau and Ring [20]. Our model gives the agent special status in the laws of physics, which makes it somewhat awkward to analyze convergent instrumental incentives for "self preservation" or "intelligence enhancement." The existing framework allows us to model these situations, but only crudely. For example, we could design a setting where if certain regions enter certain states then the agent forever after loses all actions except for actions that have no effect, representing the "death" of the agent. Or we could create a setting where normally the agent only gets actions that have effects every hundred turns, but if it acquires certain types of resources then it can act more frequently, to model "computational resources." However, these solutions are somewhat ad hoc, and we would prefer an extension of the model that somehow modeled the agent as *part* of the environment.

It is not entirely clear how to extend the model in such a fashion at this time, but the "space-time embedded intelligence" model of Orseau and Ring [20] and the "reflective oracle" framework of Fallenstein, Taylor, and Christiano [21] both offer plausible starting points. Using the latter framework, designed to analyze complicated environments which contain powerful agents that reason about the environment that contains them, might also lend some insight into how to further extend our model to give a more clear account of how the agent handles situations where other similarly powerful agents exist and compete over resources. Our existing model can handle multi-agent scenarios only insofar as we assume that the agent has general-purpose methods for predicting the outcome of its actions in various regions, regardless of whether those regions also happen to contain other agents.

6.2 Conclusions

Our model is a simple one, but it can be used to validate Omohundro's intuitions about "basic AI drives" [5], and Bostrom's "instrumental convergence thesis" [4]. This suggests that, in the long term, by default, powerful AI systems are likely to have incentives to self-preserve and amass resources, even if they are given seemingly benign goals. If we want to avoid designing systems that pursue anti-social instrumental incentives, we will have to design AI systems carefully, especially as they become more autonomous and capable.

The key question, then, is one of designing principled methods for robustly removing convergent instrumental incentives from an agent's goal system. Can we design a highly capable autonomous machine that pursues a simple goal (such as curing cancer) without giving it any incentives to amass resources, or to resist modification by its operators? If yes, how? And if not, what sort of systems might we be able to build instead, such that we could become confident they would not have dangerous effects on the surrounding environment as it pursued its goals?

This is a question worth considering well before it becomes feasible to create superintelligent machines in the sense of Bostrom [3], because it is a question about

what target the field of artificial intelligence is aiming towards. Are we aiming to design powerful autonomous agents that maximize some specific goal function, in hopes that this has a beneficial effect on the world? Are we aiming to design powerful tools with such limited autonomy and domain of action that we never need to worry about the systems pursuing dangerous instrumental subgoals? Understanding what sorts of systems can avert convergent instrumental incentives in principle seems important before we can begin to answer this question.

Armstrong [22] and Soares et al. [23] have done some initial study into the design of goals which robustly avert certain convergent instrumental incentives. Others have suggested designing different types of machines, which avoid the problems by pursuing some sort of "limited optimization." The first suggestion of this form, perhaps, came from Simon [24], who suggested designing agents that "satisfice" expected utility rather than maximizing it, executing any plan that passes a certain utility threshold. It is not clear that this would result in a safe system (after all, building a powerful consequentialist sub-agent is a surefire way to satisfice), but the idea of pursuing more "domestic" agent architectures seems promising. Armstrong, Sandberg, and Bostrom [16] and Taylor [17] have explored a few alternative frameworks for limited optimizers.

Though some preliminary work is underway, it is not yet at all clear how to design AI systems that reliably and knowably avert convergent instrumental incentives. Given Omohundro's original claim [5] and the simple formulations developed in this paper, though, one thing is clear: powerful AI systems will not avert convergent instrumental incentives by default. If the AI community is going to build powerful autonomous systems that reliably have a beneficial impact, then it seems quite prudent to develop a better understanding of how convergent instrumental incentives can be either averted or harnessed, sooner rather than later.

Acknowledgements

The core idea for the formal model in this paper is due to Benya Fallenstein. We thank Rob Bensinger for notes and corrections. We also thank Steve Rayhawk and Sam Eisenstat for conversations and comments. This work is supported by the Machine Intelligence Research Institute.

References

[1] John Brockman, ed. *What to Think About Machines That Think. Today's Leading Thinkers on the Age of Machine Intelligence*. HarperCollins, 2015.

[2] Stuart J. Russell and Peter Norvig. *Artificial Intelligence. A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2010.

[3] Nick Bostrom. *Superintelligence. Paths, Dangers, Strategies*. New York: Oxford University Press, 2014.

[4] Nick Bostrom. "The Superintelligent Will. Motivation and Instrumental Rationality in Advanced Artificial Agents". In: *Minds and Machines* 22.2 (2012): *Theory and Philosophy of AI*. Ed. by Vincent C. Müller. special issue, pp. 71–85. DOI: 10.1007/s11023-012-9281-3.

[5] Stephen M. Omohundro. "The Basic AI Drives". In: *Artificial General Intelligence 2008. Proceedings of the First AGI Conference*. Ed. by Pei Wang, Ben Goertzel, and Stan Franklin. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS, 2008, pp. 483–492.

[6] Eliezer Yudkowsky. *Complex Value Systems are Required to Realize Valuable Futures*. The Singularity Institute, San Francisco, CA, 2011. URL: http://intelligence.org/files/ComplexValues.pdf.

[7] Nate Soares. *The Value Learning Problem*. Tech. rep. 2015–4. Berkeley, CA: Machine Intelligence Research Institute, 2015. URL: https://intelligence.org/files/ValueLearningProblem.pdf.

[8] Mark R Waser. "Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence." In: *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures*. Melno Park, CA: AAAI Press, 2008, pp. 195–200.

[9] Francesco Albert Bosco Cortese. "The Maximally Distributed Intelligence Explosion". In: *AAAI Spring Symposium Series*. AAAI Publications, 2014.

[10] Nate Soares and Benja Fallenstein. *Questions of Reasoning Under Logical Uncertainty*. Tech. rep. 2015–1. Berkeley, CA: Machine Intelligence Research Institute, 2015. URL: https://intelligence.org/files/QuestionsLogicalUncertainty.pdf.

[11] John Storrs Hall. *Beyond AI. Creating the Conscience of the Machine*. Amherst, NY: Prometheus Books, 2007.

[12] Frank Tipler. "If You Can't Beat 'em, Join 'em". In: *What to Think About Machines That Think. Today's Leading Thinkers on the Age of Machine Intelligence*. Ed. by John Brockman. HarperCollins, 2015, pp. 17–18.

[13] Steven Pinker. "Thinking Does Not Imply Subjugating". In: *What to Think About Machines That Think. Today's Leading Thinkers on the Age of Machine Intelligence*. Ed. by John Brockman. HarperCollins, 2015, pp. 5–8.

[14] Mark Pagel. "They'll Do More Good Than Harm". In: *What to Think About Machines That Think. Today's Leading Thinkers on the Age of Machine Intelligence*. Ed. by John Brockman. HarperCollins, 2015, pp. 145–147.

[15] Robert A. Freitas Jr. *Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations*. Foresight Institute. Apr. 2000. URL: http://www.foresight.org/nano/Ecophagy.html (visited on 07/28/2013).

[16] Stuart Armstrong, Anders Sandberg, and Nick Bostrom. "Thinking Inside the Box. Controlling and Using an Oracle AI". In: *Minds and Machines* 22.4 (2012), pp. 299–324. DOI: 10.1007/s11023-012-9282-2.

[17] Jessica Taylor. "Quantilizers: A Safer Alternative to Maximizers for Limited Optimization". In: (forthcoming). Submitted to AAAI 2016.

[18] Gerd Gigerenzer and Reinhard Selten, eds. *Bounded Rationality. The Adaptive Toolbox*. Dahlem Workshop Reports. Cambridge, MA: MIT Press, 2001.

[19] Shane Legg and Marcus Hutter. "A Universal Measure of Intelligence for Artificial Agents". In: *IJCAI-05. Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30–August 5, 2005*. Ed. by Leslie Pack Kaelbling and Alessandro Saffiotti. Lawrence Erlbaum, 2005, pp. 1509–1510. URL: http://www.ijcai.org/papers/post-0042.pdf.

[20] Laurent Orseau and Mark Ring. "Space-Time Embedded Intelligence". In: *Artificial General Intelligence. 5th International Conference, AGI 2012, Oxford, UK, December 8–11, 2012. Proceedings*. Lecture Notes in Artificial Intelligence 7716. New York: Springer, 2012, pp. 209–218. DOI: 10.1007/978-3-642-35506-6_22.

[21] Benja Fallenstein, Jessica Taylor, and Paul F. Christiano. "Reflective Oracles: A Foundation for Game Theory in Artificial Intelligence". In: *Logic, Rationality, and Interaction. Fifth International Workshop, LORI-V, Taipei, Taiwan, October 28–31, 2015. Proceedings*. Ed. by Wiebe van der Hoek, Wesley H. Holliday, and Wen-fang Wang. FoLLI Publications on Logic, Language and Information. Springer, forthcoming.

[22] Stuart Armstrong. *Utility Indifference*. 2010-1. Oxford: Future of Humanity Institute, University of Oxford, 2010. URL: http://www.fhi.ox.ac.uk/utility-indifference.pdf.

[23] Nate Soares et al. "Corrigibility". Paper presented at the 1st International Workshop on AI and Ethics, held within the 29th AAAI Conference on Artificial Intelligence (AAAI-2015). Austin, TX, 2015. URL: http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124.

[24] Herbert A. Simon. "Rational Choice and the Structure of the Environment". In: *Psychological Review* 63.2 (1956), pp. 129–138.