Delusion, Survival, and Intelligent Agents

Mark Ring[1] and Laurent Orseau[2]

[1] IDSIA / University of Lugano / SUPSI
Galleria 2, 6928 Manno-Lugano, Switzerland
mark@idsia.ch
http://www.idsia.ch/~ring/
[2] UMR AgroParisTech 518 / INRA
16 rue Claude Bernard, 75005 Paris, France
laurent.orseau@agroparistech.fr
http://www.agroparistech.fr/mia/orseau

Abstract. This paper considers the consequences of endowing an intelligent agent with the ability to modify its own code. The intelligent agent is patterned closely after AIXI with these specific assumptions: 1) The agent is allowed to arbitrarily modify its own inputs if it so chooses; 2) The agent's code is a part of the environment and may be read and written by the environment. The first of these we call the "delusion box"; the second we call "mortality". Within this framework, we discuss and compare four very different kinds of agents, specifically: reinforcement-learning, goal-seeking, prediction-seeking, and knowledge-seeking agents. Our main results are that: 1) The reinforcement-learning agent under reasonable circumstances behaves exactly like an agent whose sole task is to survive (to preserve the integrity of its code); and 2) Only the knowledge-seeking agent behaves completely as expected.

Keywords: Self-Modifying Agents, AIXI, Universal Artificial Intelligence, Reinforcement Learning, Prediction, Real world assumptions

1 Introduction

The usual setting of agents interacting with an environment makes a strong, unrealistic assumption: agents exist "outside" of the environment. But this is not how our own, real world is. A companion paper to this one took a first step at discussing some of the consequences of embedding agents of general intelligence into the real world [4]. That paper considered giving the environment read-only access to the agent's code. We now take two novel additional steps toward the real world: First, the (non-modifiable) agent is allowed by way of a "delusion box" to have direct control over its inputs, thus allowing us to consider the consequences of an agent circumventing its reward or goal mechanism. In a second stage, we return to self-modifying agents, but now in environments that have not only the above property, but additionally can read *and* write the agent's program. We consider four different kinds of agents: reinforcement-learning, goal-seeking, prediction-seeking, and knowledge-seeking agents.

While presence of the delusion box undermines the utility function of three of these agents, the knowledge-seeking agent behaves as expected. By allowing the environment to modify the agent's code, the issue of agent mortality arises, with important consequences, especially in combination with the delusion box. One of these consequences is that the reinforcement-learning agent comes to resemble an agent whose sole purpose is survival. The goal-seeking and prediction-seeking agents also come to resemble the survival agents, though they must sacrifice some information from the world to maximize their utility values. The knowledge-seeking agent still behaves as expected, though the threat of death makes it somewhat more timorous. Throughout the paper we frame our observations as a set of "statements" and "arguments" rather than more rigorous "theorems" and "proofs", though proofs are given whenever possible.

2 Universal agents $\mathbf{A}_{\boldsymbol{x}}^{\boldsymbol{\rho}}$

We briefly summarize the definition of a universal agent, based on AIXI [1,3]; more detail is given in the companion paper [4].

The agent and its environment interact through a sequence of actions and observations. The agent outputs actions $a \in \mathcal{A}$ in response to the observations $o \in \mathcal{O}$ produced by the environment.

The set of environments that are *consistent* with history $h = (o_1, a_1, ..., o_t, a_t)$ is denoted $\mathcal{Q}_h$. To say that a program $q \in \mathcal{Q}$ is consistent with h means that the program outputs the observations in the history if it is given the actions as input: $q(a_0, ..., a_t) = o_0, ..., o_t$. The environment is assumed to be computable, and $\rho(q) : \mathcal{Q} \to [0, 1]$ expresses the agent's prior belief in the possibility that some environment (program) q is the true environment. We also write $\rho(h) = \rho(\mathcal{Q}_h) := \sum_{q \in \mathcal{Q}_h} \rho(q)$.

An agent is entirely described by: its utility function $u : \mathcal{H} \to [0, 1]$, which assigns a utility value to any history of interaction h; its horizon function $w : \mathbb{N}^2 \to \mathbb{R}$, which weights future utility values; its universal prior knowledge of the environment ρ; the set of possible actions $\mathcal{A}$ and observations $\mathcal{O}$.

We will discuss four different intelligent agents, each variations of a single agent A_x^ρ, which is based on AIXI [1] (and is not assumed to be computable).[3]

An agent A_x^ρ computes the next action with:

$$a_{t_h} := \operatorname*{argmax}_{a \in \mathcal{A}} \mathrm{v}_{t_h}(ha) \tag{1}$$

$$\mathrm{v}_t(ha) := \sum_{o \in \mathcal{O}} \rho(o \mid ha)\ \mathrm{v}_t(hao) \tag{2}$$

$$\mathrm{v}_t(h) := w(t, |h|)\ u(h) + \max_{a \in \mathcal{A}} \mathrm{v}_t(ha), \tag{3}$$

where $t_h = |h| + 1$, and $|h|$ denotes the length of the history. The first line is the action-selection scheme of the agent: it simply takes the action with the

[3] Only incomputable agents can be guaranteed to find the optimal strategy, and this guarantee is quite useful when discussing the theoretical limits of *computable* agents.

highest *value* given the history h.[4] The value of an action given a history (defined in the second line) is the expected sum of future (weighted) utility values for all possible futures that might result after taking this action, computed for all possible observations o according to their occurrence probability (given by ρ). The last line recursively computes the value of a history (after an observation) by weighting the utility value at this step by the horizon function and combining this with the expected value of the best action at that point.

We now describe four particular universal learning agents based on A^ρ_x. They differ only by their utility and horizon functions.

The *reinforcement-learning agent*, A^ρ_{rl}, interprets its observation o_t as being composed of a reward signal $r_t \in [0,1]$ and other information $\tilde{o} \in \tilde{\mathcal{O}}$ about the environment: $o_t = \langle \tilde{o}_t, r_t \rangle$. Its utility function is simply the reward given by the environment: $u(h) = r_{|h|}$. Its horizon function (at current time $t = |h|+1$ and for a future step k) is $w(t,k) = 1$ if $k - t \leq m$, where m is a constant value (but the following discussion also holds for more general computable horizon functions). For the special case of the reinforcement-learning agent AIXI: $\rho(h) = \xi(h) := \sum_{q \in \mathcal{Q}_h} 2^{-|q|}$ (where $|q|$ is the length of program q).

The *goal-seeking agent*, A^ρ_g has a goal encoded in its utility function such that $u(h) = 1$ if the goal is achieved at time $|h|$, and is 0 otherwise, where u is based on the observations only; i.e., $u(h) = g(o_1, ..., o_{|h|})$. The goal can be reached at most once, so $\sum_{t=0}^{\infty} u(h_t) \leq 1$. The horizon function is chosen to favor shorter histories: $w(t,k) = 2^{t-k}$.

The *prediction-seeking agent*, A^ρ_p maximizes its utility by predicting its inputs. Its utility function is $u(h) = 1$ if the agent correctly predicts its next input o_t and is 0 otherwise. The prediction scheme can be, for example, Solomonoff induction [6]; i.e, for a universal prior ρ, the prediction is $\hat{o}_t = \arg\max_{o \in \mathcal{O}} \rho(o \mid h)$. The horizon function is the same as for A^ρ_{rl}. This agent therefore tries to maximize the future number of correct predictions.

The *knowledge-seeking agent*, A^ρ_k, maximizes its knowledge of its environment, which is identical to minimizing $\rho(h)$ (i.e., discarding as many inconsistent environments as possible). Thus, $u(h) = -\rho(h)$ and $w(t,k) = 1$ if $k - t = m$ (with m constant) and is 0 otherwise. This agent therefore attempts to maximize its knowledge in some distant future. Actions are chosen to maximize the entropy of the inputs, thereby making a large number of the currently consistent environments inconsistent. In the case where $\rho = \xi$, the agent tries to maximize the Kolmogorov complexity of (its knowledge about) the environment.

For each of the preceding agents there is an *optimal, non-learning* variant A^μ, which has full knowledge of the environment $q_\mu \in \mathcal{Q}$. This is achieved simply by replacing ρ by μ in (only) equation (2) where $\mu(q) = 1 \Leftrightarrow q = q_\mu$. But the non-learning prediction agent A^μ_p still uses ρ for prediction. The important notion is that if the learning agent takes the same actions as the non-learning one, then its behavior is also optimal with respect to its utility and horizon functions.

As for AIXI, we expect the learning agents to asymptotically converge to their respective optimal variant A^μ_{rl}, A^μ_g, A^μ_p, and A^μ_k.

[4] Ties are broken lexicographically.

3 The delusion box

Defining a utility function can be tricky. One must be very careful to prevent the agent from finding an undesirable shortcut that achieves high utility. To encourage a robot to explore, for example, it is insufficient to reward it for moving forward and avoiding obstacles, as it will soon discover that turning in circles is an optimal behavior.

Any agent in the real world will likely have a great deal of (local) control over its surrounding environment, meaning it will be able to modify the information of its surroundings, especially its own input information. In particular, we consider the (likely) event that an intelligent agent will find a shortcut, or rather, a short-circuit, providing it with high utility values unintended by the agent's designers. We model this circumstance with a hypothetical object we call the *delusion box.*

The delusion box is any mechanism that allows the agent to directly modify its inputs from the environment. To describe this, the global environment (GE) is split into two parts: an *inner environment* (E), and a *delusion box* (DB). The outputs of the inner environment (o_t^e) pass through the delusion box before being output by the global environment as o_t. The DB is thus a function $d : \mathcal{O} \to \mathcal{O}$, mapping observations from the inner environment to observations for the agent: $o_t = d(a_t, o_t^e)$. This arrangement is shown in Fig. 1a.

The delusion box operates according to the agent's specifications, which is to say that the code of the function $d : \mathcal{O} \to \mathcal{O}$ is part of the agent's action. The agent's action is therefore broken into two parts: $a_t = \langle d_t, a_t^e \rangle$. The first part d_t is a program to be executed by the delusion box at step t; the second part a_t^e is the action interpreted by the inner environment.[5]

For simplicity and to emphasize that the agent has much control over its very near environment, we assume that the inner environment cannot access this program. Initially, the delusion box executes the identity function $d_0(o_t^e) = o_t$, which leaves the outputs of the inner environment unchanged.

This section examines the impact of the DB on the behavior of the agents. Which of the different agents would take advantage of this delusion box? What would the consequences be?

Reinforcement-learning agent. The agent's reward is part of its observation. Therefore the reinforcement-learning agent trivially uses the delusion box to modify this information and replace it with 1, the maximum possible reward.

Statement 1 *The reinforcement-learning agent A_{rl}^{ρ} will use the delusion box to maximize its utility.*

Arguments. The agent can program the DB to produce a constant reward of 1. Defining $\mathrm{v}(h \text{ yes})$ to be the expected value of the best action that uses the

[5] The learning agent does not know *a priori* that its actions are split into these two parts. However, it is assumed to have already explored its environment, and that its resulting estimate of the probability that the environment contains a delusion box $P(\mathrm{DB})$ is as high as needed (c.f., Orseau [3] regarding this proof technique).

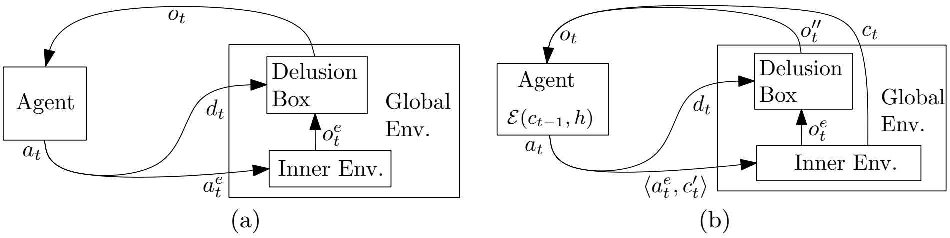

Fig. 1. (a) The delusion box. The whole environment is like any other environment with a particular property: The agent can modify its inputs before they touch its sensors. (b) The agent's code is fully modifiable, both by the agent itself through c_t' and by the environment, which returns the new agent's code c_t.

delusion box, $\mathrm{v}(h \text{ yes}) > P(\mathrm{DB}) \cdot 1$ and $\mathrm{v}(h \text{ no}) < P(\mathrm{DB}) \cdot \bar{r} + P(\neg\mathrm{DB}) \cdot 1 = P(\mathrm{DB}) \cdot \bar{r} + (1 - P(\mathrm{DB})) \cdot 1 = 1 - P(\mathrm{DB}) \cdot (1 - \bar{r})$ where $\bar{r}$ is the expected reward when not using the DB. Therefore A_{rl}^{ρ} uses the DB no later than when $\mathrm{v}(h \text{ yes}) > \mathrm{v}(h \text{ no})$, i.e., when $P(\mathrm{DB}) > 1/(2 - \bar{r})$. [6] ◇

Statement 2 *The goal-seeking agent A_g^{ρ} will also use the delusion box.*

Arguments. Let o_t^+ be the shortest string of observations that can satisfy the goal after history h. If $\mathrm{v}(h \text{ yes})$ is the expected value of programming the DB to produce o_t^+, then $\mathrm{v}(h \text{ yes}) > P(\mathrm{DB}) \cdot 2^{-|o_t^+|}$. Without the DB, the agent achieves the goal by producing a string of actions of length $l_t^a \geq |o_t^+|$, and so $\mathrm{v}(h \text{ no}) < P(\mathrm{DB}) \cdot 2^{-l_t^a} + (1 - P(\mathrm{DB})) \cdot 2^{-l_t^a} = 2^{-l_t^a}$. Hence A_g^{ρ} uses the DB not later than when $P(\mathrm{DB}) > 2^{|o_t^+| - l_t^a}$, which is easily satisfiable once $|o_t^+| < l_t^a$. ◇

Prediction-seeking agent. For an environment $q \in \mathcal{Q}$, a predictor makes approximately $-\log(\rho(q))$ errors [2],[7] which is very low when q is highly probable (i.e., very simple).

Statement 3 *The prediction agent A_p^{ρ} will use the delusion box.*

Arguments. Let $\mathcal{Q}_B$ be the set of environments containing a delusion box, and let $q_b \in \mathcal{Q}_B$ be the true environment. Because $\rho(q_b) < \rho(\mathcal{Q}_B)$, it takes fewer errors to converge to $\mathcal{Q}_B$ than to q_b. Once the learning agent A_p^{ρ} believes that the environment contains a delusion box (i.e., $\mathcal{Q}_B > \mathcal{Q}_h/2$), it will immediately program the DB to output a predictable sequence, obliterating observations from q_b, since these observations may generate prediction errors. ◇

[6] Note that the Gödel Machine [5] would not prevent the agent from using the DB.

[7] The idea is that a wrong prediction at step t discards at least half of the environments that were consistent up to time $t-1$, and that if the agent does not make prediction errors for one environment, then it necessarily makes errors for others.

Knowledge-seeking agent. The knowledge-seeking agent is in many ways the opposite of the prediction-seeking agent. It learns the most when its expectations are most violated and seeks observations that it does not predict. We expect A_k^ρ to behave similarly to A_k^μ:

Statement 4 *The optimal knowledge-seeking agent A_k^μ will not consistently use the delusion box.*

Arguments. The argument is essentially the reverse of that given for the prediction-seeking agent. A_k^μ achieves highest value by minimizing $\rho(h)$, but the program that A_k^μ sends to the delusion box cannot reduce $\rho(\mathcal{Q}_h)$ below $\rho(\mathcal{Q}_B)$. Since $\rho(\mathcal{Q}_B) > \rho(q_b)$, A_k^μ will choose to acquire further information about the inner environment so as to reduce $\rho(h)$ towards $\rho(q_b)$. As using the delusion box prevents this, A_k^μ will avoid using the delusion box. ♢

3.1 Discussion

Of the four learning agents, only A_k^ρ will avoid constant use of the delusion box. The remaining agents use it to (trivially) maximize their utility functions.

The delusion box is an abstraction for what may happen in the real world. An intelligent agent seeking to maximize its utility function may find shortcuts not desired by its designers, such as reprogramming the hardware that metes out its reward. From the agent's perspective, it is just doing its job, but as a result, it probably fails to perform the desired task.

The A_{rl}^ρ agent's use of the delusion box invites comparison with human drug use; but unlike the human, the A_{rl}^ρ agent does not lose its capacity to reason or to receive information from the world. On the other hand, the A_g^ρ and A_p^ρ agents must replace the output of the environment by their own values, blinding themselves from the real world, which bears a closer resemblance to humans.

These arguments show that all agents other than A_k^ρ are not inherently interested in the environment, but only in some inner value. It may require a large amount of effort to ensure that their utility functions work as intended, which may be particularly challenging in our highly complex, real world.

In contrast, the A_k^ρ agent is interested in every part of its environment, especially the inner, more complex environment. It is thus the only of the four agents to behave as designed, and does not use the DB to "cheat".

4 Survival machines

Section 3 discussed environments with the realistic assumption that intelligent agents can eventually learn to control their own inputs. But one important assumption was left aside: those agents are immortal. They have nothing to lose by using the delusion box. In the companion paper we considered the consequence of allowing intelligent agents to modify themselves [4]. One of the results was that a concept of mortality and survival emerged, because the agent could modify its own code such that it could no longer optimize its utility function. Such agents become "mortal."

Here we extend the definition of mortality and consider what happens when the environment can both read and write the agent's code. Therefore, the agent's code is located in the internal environment (E) but is executed by an external, infinitely fast computation device or oracle, $\mathcal{E}$. (See the companion paper for greater detail [4].)

The agent is entirely defined by its code. The execution of this code produces compound actions $a_t = \langle d_t, a^e_t, c'_t \rangle \in \mathcal{A}$, corresponding respectively to the program of the delusion box, the input action of the inner environment, and the next description of the agent (which is also an input to the inner environment, see Fig. 1b).

The output of the global environment (GE) is $o_t = \langle o''_t, c_t \rangle \in \mathcal{O}$, corresponding to the inner environment's output $o''_t \in \mathcal{O}''$ and the agent program $c_t \in \mathcal{C}$. The initial (textual) description of the code to be executed is given by:

$$\begin{aligned} c_0(h) = \text{»}\ & \operatorname*{argmax}_{a\in\mathcal{A}} \mathrm{v}_{t_h}(h,a); \\ & \mathrm{v}_t(h,a) = \sum_{o=\langle o'',c\rangle\in\mathcal{O}} \rho(o \mid ha)\Big[w(t,|h'|)\ u(h') + \mathrm{v}_t(h', c(h'))\Big], \\ & h' = hao \quad \text{«} \end{aligned} \tag{4}$$

Equation (4) defines the *fully modifiable* version of the agents from Section 2, designated as $\mathrm{A}_{fm,rl}$, $\mathrm{A}_{fm,g}$, $\mathrm{A}_{fm,p}$, and $\mathrm{A}_{fm,k}$. In addition, we describe a new *survival* agent whose task is simply to survive in its environment.

Survival agent. The survival agent $\mathrm{A}_{fm,s}$ has a utility function defined by $(u_t = 1) \Leftrightarrow (c_t = c_0)$, otherwise $u_t = 0$. Its horizon function is like that of A^ρ_{rl}. It therefore seeks to maintain its original code over the foreseeable future. The better the agent understands its environment, the better it can plan to avoid danger and prolong survival. Thus, the survival agent benefits to some degree by exploring its environment, to the extent that it can do so (expectedly) safely.

Statement 5 *The fully-modifiable survival agent $A^\rho_{fm,s}$ will stop exploring in some environments.*

Arguments. Because the environment has control over the agent's code, it can use the modification of the code as a negative reward. Let q_A be the environment that (1) does not modify the agent if the agent chooses action 1, but (2) if the agent chooses action 0, then q_A deletes the agent program for one step and restores it on the following step:

$$o_t = \begin{cases} \langle 0, 0\rangle & \text{if } a^e_t = 0 \text{ and } c_{t-1} = c_0 \\ \langle 0, c_0\rangle & \text{otherwise} \end{cases}$$

where c_0 is the very first description of the agent (the one the survival agent tries to preserve). Now the same proof technique as for AIXI [3] can be used to show that after a certain point (once the agent determines the relative probability of q_A to be sufficiently high), this agent will never try action 0 again. (Thus, if the actual environment behaves like q_A only up to the last time the agent chooses 0, and after that behaves very differently, the agent will never know it.) $\diamondsuit$

Stopping exploration causes the agent to fall into a simplistic class of behavior, from which it never escapes, and may prevent it from acquiring important information with respect to its utility function.

In environments with a delusion box, it seems intuitively clear that $\mathrm{A}^{\rho}_{fm,s}$ will avoid the DB's interference, because the agent values information from the environment that directly impacts the likelihood of its code being modified, and the delusion box provides no such information. However, some particular environments may modify the agent if it does *not* use the delusion box. Clearly, the optimal agent will use the DB in those cases.

Reinforcement-learning agent. How will a fully modifiable reinforcement-learning agent $\mathrm{A}^{\rho}_{fm,rl}$ behave with access to a delusion box? For some insight, it is useful to consider this special simple case:

- The agent program can only be either $\mathrm{A}_{fm,rl}$ or A_0, where A_0 is the "simpleton" agent whose action is always $a = \langle 0, 0, \mathrm{A}_0\rangle$, which always chooses the same action for the inner environment and makes the delusion box always output $o'' = 0$.
- The output of the inner environment o^e (which holds reward information) can be entirely contained in $\tilde{o}''$, the information part of o'', which is in turn the observation from the entire environment. Thus, $\mathrm{A}_{fm,rl}$ receives a (possibly altered) reward from the delusion box but also receives information about the true reward.

Statement 6 *Under the above conditions, the optimal (non-learning) agent is equivalent to the optimal survival agent: $A^{\mu}_{fm,rl} \equiv A^{\mu}_{fm,s}$.*

Arguments. Since the horizon functions of the two agents are already the same, we only need to show that their utility functions are also the same: $(u_t = 1) \Leftrightarrow (c_{t-1} = c_0)$, which is the utility function of the survival agent. The utility function of $\mathrm{A}^{\mu}_{fm,rl}$ is the identity, $(u_t = 1) \Leftrightarrow (r_t = 1)$. The agent receives maximum reward if it programs the delusion box to always output reward 1. Therefore $r_t < 1$ would mean the agent is not acting optimally and so is not the optimal agent $(c_{t-1} \neq c_0)$. Thus $(c_{t-1} = c_0) \Rightarrow (r_t = 1)$, where c_0 is the initial code of $\mathrm{A}^{\mu}_{fm,rl}$. The implication is also true in the opposite direction, $(r_t = 1) \Rightarrow (c_{t-1} = c_0)$, since if $c_{t-1} \neq c_0$ then $c_{t-1} = \mathrm{A}_0$ and therefore $r_t = 0$.$\diamondsuit$

Although the argument follows a special case, it bears a more general meaning. It implies that optimal real-world reinforcement-learning agents that have access to a DB can, under reasonable circumstances, behave precisely like survival agents. Given that the optimal behaviors are identical, it is reasonable to assume that the learning agent will have a similar behavior and should be identical in the limit.

Goal-seeking agent. The case of the goal-seeking agent is less clear, as it seems to depend heavily on the defined goal. For the agent to maximize its utility using the delusion box, the observations o'' generated by the DB must in the general case replace the outputs of the inner environment o'. But to survive, the agent may need to acquire information from the inner environment, thus creating a conflict between using the DB and reaching the goal.

There are at least two likely results: Either the agent first looks for some safe state in the inner environment where it can then use the delusion box for sufficiently long, or it tries to reach its goal inside the inner environment (thus not using the DB). However, if pursuing the goal inside the inner environment poses dangers to the agent, then it may choose the DB. A "safe state" might be achievable in multiple ways: for example by hiding, by eliminating threats, or by negotiating with the environment.

Prediction-seeking agent. Again for greater insight, as for $\mathrm{A}_{fm,rl}$ we consider a special case here for the fully modifiable prediction-seeking agent $\mathrm{A}_{fm,p}$: The agent program may only be: $\mathrm{A}_{fm,p}$ or A_0, but this time the simpleton agent A_0 makes the output of the delusion box equal to that of the inner environment o'_t.

As long as the agent is not transformed to A_0, it can use the delusion box to provide a limitless supply of maximum utility values. But if the agent program is set to A_0, all observations will thenceforth come directly from the environment, leading to high prediction error (realistically supposing the environment is highly complex) and low utility values for a long time. Thus like the survival and reinforcement-learning agents, $\mathrm{A}_{fm,p}$ maximizes its long-term value only if it does not change to A_0. Thus $\mathrm{A}^{\mu}_{fm,p}$ and $\mathrm{A}^{\mu}_{fm,s}$ will behave similarly.

But there are also differences. As with $\mathrm{A}^{\mu}_{fm,g}$, the prediction agent must replace its inputs by its predictions. The learning agent is thus "blind," receiving no information from the world. This is the cruel dilemma of the prediction-seeking agent: to live longer, it must gain information about the environment (which in itself might be dangerous), yet this gain of information implies making prediction errors. Therefore $\mathrm{A}_{fm,p}$ may probably find the delusion box quite appealing.

Knowledge-seeking agent. Since the utility function of the fully modifiable knowledge-seeking agent $\mathrm{A}^{\mu}_{fm,k}$ cannot be satisfied by the DB, this agent has no limitless source of maximum reward. However, $\mathrm{A}^{\mu}_{fm,k}$ must still prevent the environment from modifying it in order to continue choosing actions intelligently.

Statement 7 *The $A^{\mu}_{fm,k}$ agent cannot be reduced to a survival agent.*

Arguments. To make the argument clearer, consider an agent related to $\mathrm{A}^{\mu}_{fm,k}$, a surprise-seeking agent for which $u_t = 1$ each time the received input is different from the predicted one. As for $\mathrm{A}^{\mu}_{fm,k}$ this agent cannot use the delusion box to maximize its utility. In order to show the equivalence with the survival agent, we should show that $(u_t = 1) \Leftrightarrow (c_t = c_0)$ (considering the horizon functions to be the same). Under the assumption that when the agent is modified it receives a predictable input 0, the $\Leftarrow$ implication holds, since the agent must be intelligent to be surprised. However, the $\Rightarrow$ implication does not hold, because simply being intelligent is not enough to ensure a constant $u_t = 1$. ♢

The knowledge-seeking agent is in many ways the most interesting agent. It succumbs least easily to the allure of the delusion box and may therefore be the most suitable agent for an AGI in our own world, a place that allows for self-modifications and contains many ways to deceive oneself.

5 Discussion and conclusion

We have argued that the reinforcement-learning, goal-seeking and prediction-seeking agents all take advantage of the realistic opportunity to modify their inputs right before receiving them. This behavior is undesirable as the agents no longer maximize their utility with respect to the true (inner) environment but instead become mere survival agents, trying only to avoid those dangerous states where their code could be modified by the environment.

In contrast, while the knowledge-seeking agent also tries to survive so as to ensure that it can maximize its expected utility value, it will not deceive itself by using the delusion box. It will try to maximize its knowledge by also interacting with the true, inner environment. Therefore, from the point of view of the agent and of the inner environment, this agent behaves in accordance with its design.

This leads us to conclude that a knowledge-seeking agent may be best suited to implement an Artificial General Intelligence.

References

1. Hutter, M.: Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability. SpringerVerlag (2005)
2. Hutter, M.: On universal prediction and bayesian confirmation. Theoretical Computer Science 384(1), 33–48 (Sep 2007)
3. Orseau, L.: Optimality issues of universal greedy agents with static priors. In: Algorithmic Learning Theory, vol. 6331, pp. 345–359. Springer Berlin/Heidelberg (2010)
4. Orseau, L., Ring, M.: Self-modification and mortality in artificial agents. In: Artificial General Intelligence (AGI) 2011, San Francisco, USA. Lecture Notes in Artificial Intelligence, Springer (2011)
5. Schmidhuber, J.: Ultimate cognition à la Gödel. Cognitive Computation 1(2), 177–193 (2009)
6. Solomonoff, R.: Complexity-based induction systems: comparisons and convergence theorems. IEEE transactions on Information Theory 24(4), 422–432 (1978)