Beyond Rewards and Values

A Non-dualistic Approach to Universal Intelligence

Akira Pyinya

akirapyinya@gmail.com

Abstract

Building an AI system that aligns with human values is believed to be a two-step process: first design a value function or learn human value using value learning methods, then maximize those values using rational agents such as AIXI agents. In order to integrate this into one step, we analyze the dualistic assumptions of AIXI, and define a new universal intelligence model that can align with human preferences or specific environments, called *Algorithmic Common Intelligence (ACI)*, which can behave the same way as examples. ACI does not have to employ rewards or value functions, but directly learns and updates hypothetical policies from experience using Solomonoff induction, while making actions according to the probability of every hypothesis. We argue that the rational agency model is a subset of ACI, and the coevolution of ACI and humans provides a pathway to AI alignment.

1. AIXI as a dualistic model

Dualistic agent and embedded agent

In most agent-based intelligence models, the agents are cleanly separated from their environments, but agents in the real world are a part of their environments. Demski and Garrabrant (2019) termed the former *dualistic agents*, and the later *embedded agents*.

A dualistic agent acts like it's playing a video game, interacts with the game only through well-defined input and output channels, such as the screen and the controller. The agent doesn't have any opportunity for self-modification. It's immortal, and its reward circuit is not vulnerable to being hacked.

On the contrary, an embedded agent is a part of the universe. There is no clear boundary between the agent and the environment. An embedded agent may improve itself, but might also modify its original goals in undesirable ways, like "directly tamping its reward circuit to get rewards in a conventional way", which was called self-rewarding (Hutter 2007) or wireheading.

It may even drop an anvil on its head to see what happens (Bensinger 2014). Those possible behaviors are not within the scope of traditional rational agency models which assume the agents are dualistic.

In order to build a new theoretic framework for embedded agents, let's investigate traditional models in search of when the dualistic assumption was introduced.

AIXI is dualistic and Solomonoff induction is not

Marcus Hutter(2003)'s AIXI is a general model for rational agents. The AIXI model consists of an agent and an environment, the agent sends out actions to the environment, while the environment sends input to the agent. Then the agent divides the input into two parts, *standard observations* and *reward inputs.*

AIXI uses Solomonoff Induction, a universal inference method (Solomonoff 1964), to assign priors to hypothetical models for the observations and rewards, so that the agent can take the actions with the highest expected reward.

The AIXI agent is dualistic: its input/output channels are well-defined, it is immortal, its goals could not be modified, and its reward signals are cleanly separated from the standard inputs (Bensinger 2014).

However, Solomonoff induction itself is not dualistic. It is supposed to "learn to correctly predict any computable sequence" (Hutter, Legg, and Vitanyi 2007), including any time series in a computable universe, with no assumption of separation between agent and environment. We can come to the conclusion that the dualistic assumption was introduced somewhere else in AIXI.

AIXI has set two barriers

Scanning through the formal statement of AIXI, we notice that two barriers were set as the dualistic assumptions of the model: the barrier between the agent and the environment, and the barrier between the standard observations and the reward inputs.

Like most, if not all agent based models, AIXI divides the universe into two parts: an *agent* and an *environment.* The agent is a system that interacts with the environment, its action (output) and perception (input) can be defined and known clearly. Modification or perishment of an agent is not allowed by the AIXI model. We call the barrier between agent and environment *the first barrier*.

After that, AIXI splits the environment input into two parts, the *standard observations* and the *reward* inputs. The agent can construct models of the world using standard observations, and the goal of the agent is set by maximizing expected rewards. Any piece of input information can be reliably categorized as either a standard observation or a reward input. We call the barrier between the standard observations and the rewards *the second barrier*.

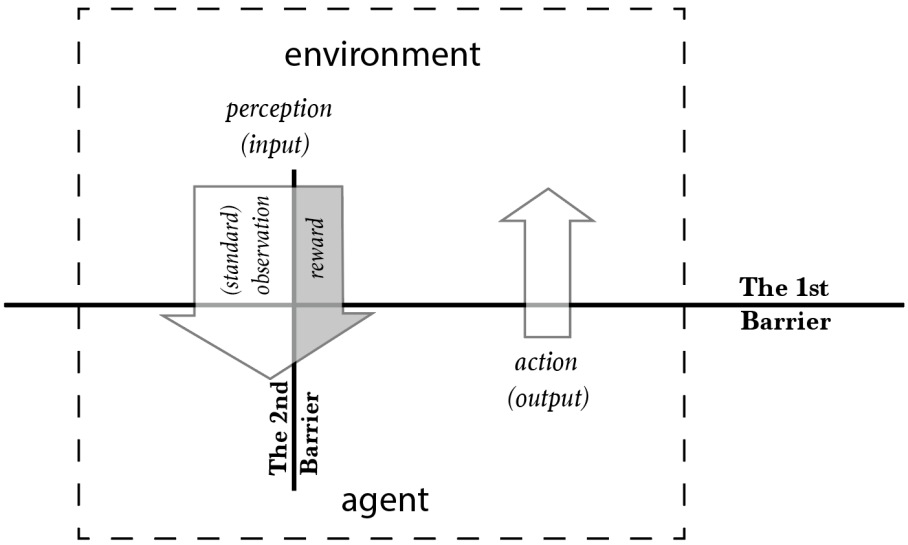


Figure 1. Two barriers in the AIXI model

2. Introducing ACI: Agents that need no reward

What can be derived if we remove the dualistic assumptions in AIXI? First, we can focus on the second barrier because it's built on the premise of the first barrier.

Without the second barrier, the environment input is no longer split into standard observations and reward inputs. As in the AIXI model, reward is interpreted as a part of the goal system (Hutter 2007), the agent will not be able to recognize its goal without the guidance of the reward input. How can an agent learn what is right and what is wrong without rewards?

Value learning

"The problem of achieving agreement between our true preference and the objective we put into the machine is called the *value alignment problem*"(Russell, Norvig 2020) . But designing the objective according to how we think a machine should behave might put the wrong values into the machine, because human value "contains considerable complexity hidden from us by our own thought processes" (Yudkowsky 2011). *Value learners* were developed to treat human goals as the goals of the agent, such as Russell (1998)'s *inverse reinforcement learning (IRL)* which learns the reward function we humans optimize with, and Dewey (2011)'s *observation-utility function* which calculates expected utility given an interaction history. Other value learning approaches include narrow value learning (Shah 2019), and natural value learning (Merwijk 2022).

But learning an explicit human value is "particularly hard because in novel situations that humans haven't seen yet, we haven't even developed the criteria by which we would evaluate" (Shah 2019). Shah proposed *narrow value learning* which has common sense and corrigibility.

However, human's preferences might change, and are not consistent between individuals. Armstrong(2018) even claims that "it is impossible to get a unique decomposition of human policy and hence get a unique human reward function".

Case-based systems

Value learning approaches broadly assume a two-step process to build a rational AI system that aligns with human values: first, get a value either by designing or by learning from human behavior, then optimize that value using knowledge learned from experience. Instead of splitting the learning process into two steps, can we attempt to directly emulate human behavior?

Currently there are a few models that directly learn policies from past experience, such as *imitation learning*, which learns action policy by imitating "right" experiences (Hussein 2017); and *case-based reasoning* (CBR), which solves problems like *common law systems*, making decisions according to previous experiences, take similar actions in similar situations (Kolodner 1992); and the *Copycat project* from Hofstadter and Mitchell (1995), who believe learning behaviors are based on analogy making.

However, these learning methods still can't work without a reward signal or utility functions, and are thought to be of limited intelligence, because they can't outperform humans who have very limited capability (Christiano 2015).

Definition of ACI

We can easily notice the similarities among imitation learning, CBR, and Solomonoff induction. All the three methods are trying to predict the next value in a data sequence. In the case of imitation learning and CBR, the "data sequence" consists of previous experiences.

Thus we can formalize imitation learning and C*BR* in the same way as the AIXI model formalizes rational agents, and arrive at the *Algorithmic Common Intelligence (ACI).*

Definition 1 (ACI) *ACI is an agent that interacts with an unknown environment in time cycles k = 1,2,3,...m. In cycle k,* x_k *is the perception(input) from the environment, and* y_k *is the action(output) of the agent. The action in the next time cycle* y_{k+1} *is the output of one policy that is chosen from a set of hypothetical policies. Each policy is a function that has a probability determined by Solomonoff induction from previous perception-action history and current perception* x_{k+1}*, which is linked to the length of Turing machine that outputs the sequence* $y_1x_2, y_2x_3,... y_{k-1}x_k$ *,* y_kx_{k+1} *.*

No reward is involved in this definition, thus *the second barrier* which divides the input into standard observation and reward input is removed.

An ACI agent consists of two parts, a *policy evaluator(PE)*, and a *Memory* . The *Memory* stores all past experiences, including perceptions from the environment and actions the agent has taken.

With all the experiences in *Memory*, *PE* evaluates all the agent's possible action policies using Solomonoff induction, calculates the probabilities of all hypothetical policies, performs an action drawn from a probability distribution of different hypotheses, then updates the probabilities of all hypothetical policies based on new data.

Just like AIXI, ACI is not computable because of the uncomputability of Solomonoff induction. Practical ACI agents should be approximations of the ACI model.

Unlike AIXI, ACI does not act deterministically. It performs actions according to the probability distributions of hypothetical policies, instead of always choosing the action with the highest probability.

However, ACI is still a partly dualistic model, because it keeps the *first barrier* which divides the universe into agents and environment. ACI agents are immortal and unchangeable in the same way as those of AIXI. We need a *General ACI (gACI)* model without the distinction between the agent and the environment, or at least without setting an explicit boundary. However, this topic is beyond the scope of this article.

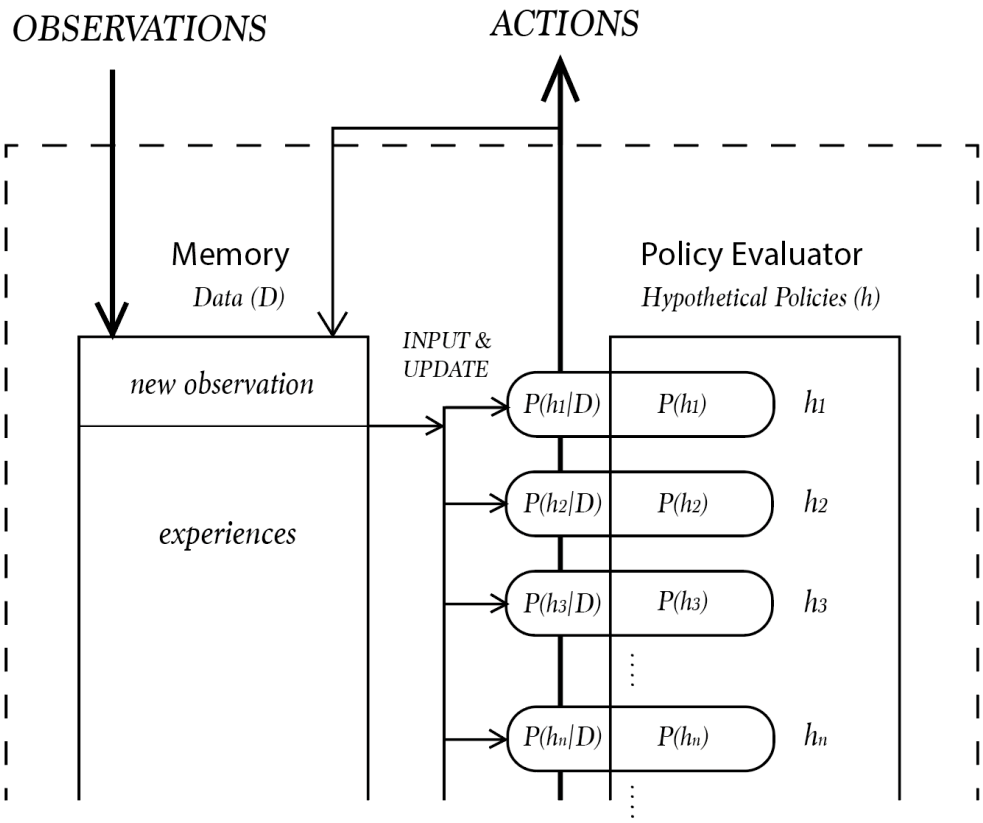


Figure 2. Structure of an ACI agent

3. Types of ACI

Natural ACI

Unlike AIXI, ACI couldn't work without the guidance from "right" experiences. But where do such "right" experiences come from?

For a natural intelligence, there is a perfect source of right experiences: its evolutionary history. All the actions of a living individual and its ancestors must be "right", at least in their own environments, otherwise this individual won't be here at all. If we have the knowledge of all those experiences, we can infer how this individual would behave. We term this model of natural intelligence *Natural ACI (nACI)*.

The behavior of an nACI agent is always right as long as it's alive, except in the scenario of human domesticating organisms, in which humans determine what behavior is right instead of organisms themselves.

However, natural intelligence in the real world could only be approximations of nACI agents, not only because ACI is uncomputable, but also because of the difficulty of storing and indexing large amounts of experience data. All the memory and hypothesis information must be saved in a lossy compressed form.

Artificial ACI

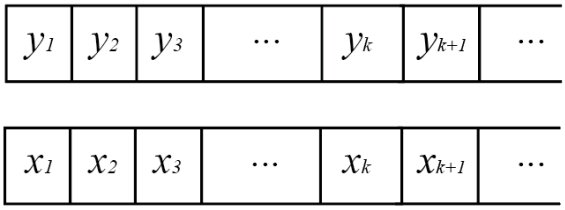
For an Artificial intelligence agent, we need abundant example data of how the agent should act in different environments. An ACI agent learns to behave the same way as the examples. We term this model of artificial intelligence *Artificial ACI (aACI)*.

Furthermore, an aACI agent needs to act autonomously after it has learned how to act from examples. We term the autonomous stage the *unsupervised stage* and the policy learning stage the *supervised stage*. However, in the unsupervised stage, the agent can still learn from the input to refine the model of the universe. Unlike the example behaviors in the supervised stage which is "right" by definition, there is no guarantee that the agent's behavior in the unsupervised stage is always right.

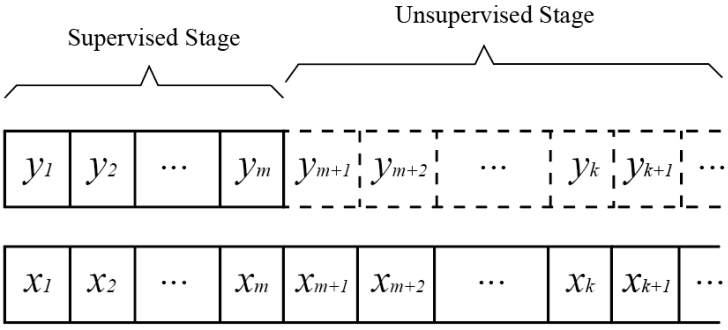
Thus we have the definition of aACI:

Definition 2 (aACI) a*ACI is an ACI agent whose action(output) sequence* $y_2, y_3, \ldots y_k$ *can only partially be used to determine the probability distribution of the agent's hypothetical policies.*

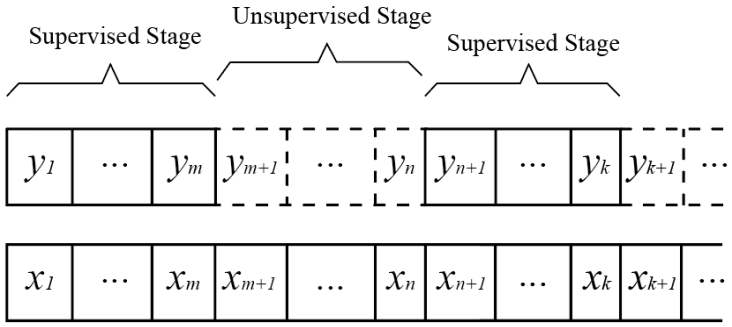
This definition does not rule out supervised stages that come after an unsupervised stage. The only constraint is that there must be a supervised stage in the beginning of the learning procedure to give some initial guidance to the agent's actions in the following stage.



(a)



(b)



(c)

Figure 3. Perception-action sequence the agent uses to determine the probability distribution of hypothetical policies.

(a) A nACI agent uses the sequence $y_1x_2, y_2x_3, \ldots y_{k-1}x_k\ , y_kx_{k+1}$

(b) An aACI agent uses the sequence $y_1x_2, y_2x_3, \ldots y_mx_m\ , \theta_1x_{m+1}\ , \ldots \theta_{k-m}x_{k+1}$,

the value of θ can be any symbol in the alphabet;

(c) An aACI agent uses the sequence $y_1x_2, y_2x_3, \ldots y_mx_m\ , \theta_1x_{m+1}\ , \ldots \theta_{n-m}x_{n+1}, y_{n+1}x_{n+2}, \ldots y_{k-1}x_k\ , y_kx_{k+1}$

Policies and sub-policies

As discussed in **Definition 1**, a policy of an agent is a function that outputs actions. Without loss of generality, we can categorize all available policy functions into:

A. *Functions that do not take precedents as variables*;
B. *Functions that take precedents as variables*;

Functions of *type A* include some boring functions, like *constant functions*, which may equal to:

```
Print (n) forever
```

Periodic functions and *monotonic functions* are also type A functions. Obviously, ACI agents that implement only type A policies can not be very intelligent.

Type B functions are more complicated, including *reflex* policies, which follow a condition-action rule, and can be written as if-then-else statements in a programming language.

Reward-oriented or *value* policies used in the AIXI model are not computable functions. But value policies used in the *AIXItl* model, a pragmatic version of AIXI which runs in limited time and limited computational resources, are a category of computable type B functions.

However, candidate policies include not only simple functions, but also their combinations, such as combinations of reflex functions and value functions. For example, for natural intelligence agents, a preference can be turned on and off by a reflex function, like the appetite would be turned off if one animal has eaten more than enough food.

We call the components of policies *sub-policies*. Inside any ACI agent, there might be up to millions of sub-functions that make up one entire policy. Here is a simple example:

```
if Lawful:
      if Self_Preservation:
            do with Trade_off(Expected_Energy_Consumption,
      Expected_Food_Gain)
```

This seems similar to a rule-based system, but the main difference between rule-based systems and ACI is that the latter implements and updates multiple hypothetical rules while the former implements only one.

In summary, each ACI agent implements multiple hypothetical policies, each policy consists of multiple sub-policies. For efficiency, sub-policies of different policies can be selected from a library of sub-policies.

Unlike the *instrumental convergence thesis* which argues that instrumental goals should converge to a final goal (Bostrom 2012), a policy in an ACI agent is not a goal for sub-policies to converge,

but an ecosystem of sub-policies that act simultaneously to reach a balance in certain environments, so that the agent performs in the same way as precedent.

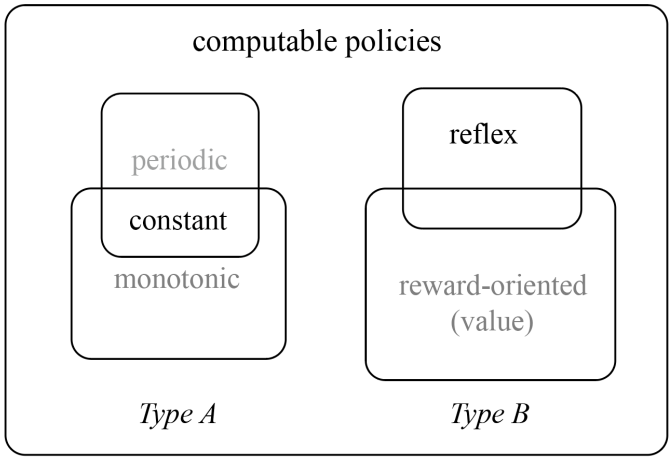


Figure 4. Types of computable policies

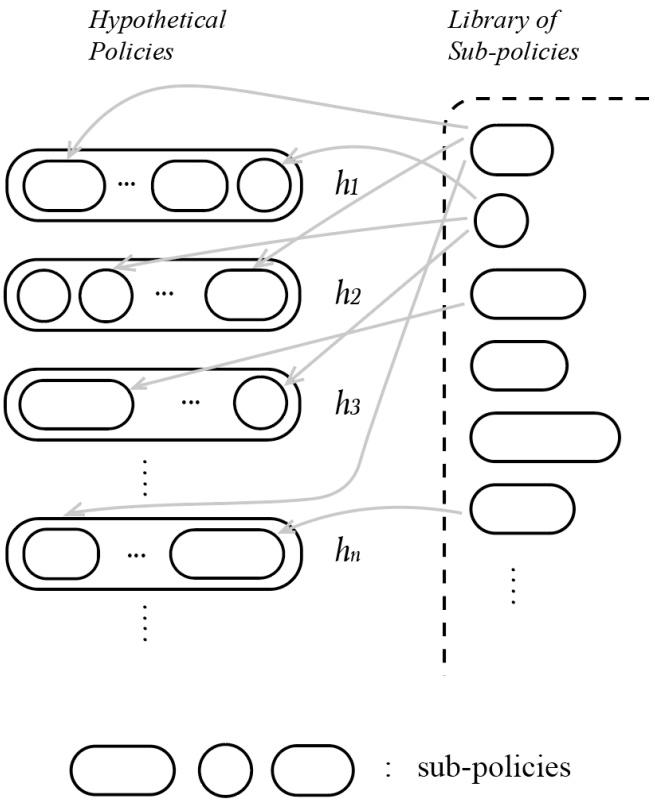


Figure 5. Hypothetical policies, sub-policies, and library of sub-policies

IRL + AIXI is a subset of aACI

Suppose there is a limited aACI agent *L*, whose candidate policies are restricted to only certain types of policies, we can term *L* as a *subset of aACI*.

The combination of IRL and AIXI is a subset of aACI. Here is the proof:

Like aACI agents, the IRL approach also tries to estimate humans' action policies, but their estimations are restricted to a special category: the value functions, which maximize certain values.

Value functions learned by a IRL process can be implemented by AIXItl agents. That's how a IRL + AIXItl combination works. This combination is equivalent to an aACI agent whose hypothetical policy range is restricted to value functions. Thus, we can come to the conclusion that IRL + AIXItl is a subset of aACI. To avoid misunderstanding, we will use *human preference* instead of human value to describe all types of policies implemented by humans.

4. An example: paperclips making aACI

The ACI model does not depict AI as a genie that can fulfill any human wishes, but a copycat that repeats its past behaviors. What can we make use of an intelligence system that can only imitate? How can a aACI agent correct itself from wrong actions that do not align with human preferences?

Here is an example of how an aACI agent makes paperclips in a reasonable way that aligns with human preferences.

Supervised stage

In the beginning there is a supervised stage, during which the agent learns from examples. The examples are generated by a human-controlled robot which makes paperclips. The perception input and action output of the robot are recorded and stored in the memory, from which the aACI agent learns the probabilities of its hypothetical policies.

The paperclip making agent is supposed to work in a world occupied by human civilization, therefore the example-generating robot should not only operate in a controlled environment well-provided with power supply and raw materials. Instead, it has to gather resources by either trading with humans or legally extracting from nature, like gaining solar power and smelting iron from sand. From such experiences, an aACI agent can learn policies made up of sub-policies such as paperclips producing, self preservation, resource acquisition, and local law abiding.

Unlike the rational agents, there's no ultimate goal or ultimate sub-policy for an aACI agent. An aACI agent is not obedient to the law only in order to make more paperclips. From provided

experiences, the agent spontaneously learns all the sub-policies and relationships between policies, such as the order of priority or possible trade-offs between sub-policies. In the case of paperclip making agents, law abiding is primary, making paperclips illegally should not be included in a "right" example, as long as the robot operator is virtuous.

Unsupervised stage

Then comes an unsupervised stage, during which the aACI agent acts autonomously with its computational power superior to human intelligence, implementing the policies it has learned in the previous stage. Compared to human operators, the intelligence agent is able to deal with more hypothetical policies and experiences, make predictions faster, and find new ways to implement old policies. In other words, the efficiency of paperclips making would increase greatly.

But in an unsupervised stage the agent does not always act correctly. If the aACI agent shares all the human experiences, it would learn policies as close to human preferences as possible, but also requires a supervised stage as long as humans' evolution history. Since any realistic agent can only share a small portion of human experience, there must be some discrepancy between policies of the agent and humans.

With the increase of agent's efficiency, some mistakes in previous policy learning processes might be easier to be noticed, some policies need to be improved, and some new policies need to be developed. Therefore, another supervised stage is required to get all those corrections done.

Another supervised stage

In this new supervised stage, the agent's actions will be judged by humans. Right actions are saved and used to improve agent's policies, while wrong actions are not. There are two causes of wrong actions:

First, the agent may have learned wrong policies which entail the same behavior as the right ones when the agent's competence is limited, but entail different behavior when its competence is superior to that of humans. For example, the agent may learn to avoid punishment instead of to obey the law, just as a few humans learned to do. Results of the two policies are almost the same until the agent gets the competence to cheat the law enforcement system.

Second, a powerful agent may change the environment in such an unprecedented way, that neither the agent nor humans have a definite answer to what behavior is beneficial in such a new environment. In the case of a paperclips maker, an aACI agent might develop nanomachine technology to produce paperclips faster, while nanomachine may change human society, and even the form of human life in an unpredictable way.

To what degree should nanomachine technology be regulated? There might be no determined answer for this question. In this supervised stage, humans must choose among the technology pathways that aACI provides, with the result that they choose their future way of life. There

should be some choices better than others, but there is no best choice, just like there is no best species in the biosphere.

Human-AI coevolution for AI alignment

In summary, there are two main causes of the ACI agent's misalignment with the human preferences: the difference between human's experiences and the agent's experiences, and different pathways that human and the agent choose in a new environment.

Different experiences might entail different preferences, but this difference could also be a reason why machines will always need humans. In the *intelligence explosion* scenario, a superintelligence machine could design an even better machine (Good 1966) , and machines can develop their own policies. But according to the ACI model, policy building requires an abundance of past experience. Even if a superintelligence has achieved a super high sampling frequency from its trillions of sensors, the information it has collected is still those within the short period after its birth, and cannot replace those collected by the lineage of *Homo sapien* in billions of years, which is stored genetically, culturally, and mentally. Just as supercomputers still need power supply from traditional industry, superintelligence still needs experience borrowed from humans or other organisms.

We need powerful AI because they create new possibilities and enrich the world. They can not only fulfill our old wishes, but also bring new environments we've never seen before, in which our preferences are not known to AI. That is why superhuman AI might be harmful. Humans need time to apprehend the new environment and develop new preferences which can be learned by the aACI agents.

In order to achieve AI alignment, we need an incremental schedule for AI improvement and humans' adaptation to live with AI, in which the progress of AI's capacity can be complemented with humans' velocity of adaptation to a new environment, and AI is able to learn humans' new preferences spontaneously without lose control. Following Lee (2020) 's book *The Coevolution*, we call this process "Human-AI Coevolution".

Could ACI be wireheaded?

As discussed above, an ACI agent can correct its policies in supervised stages. However, the learning algorithm of ACI is also vulnerable to damages. If an ACI agent always acts in a wrong way, we can conclude that the agent has not learned the right policy at all, either because the learning task is too difficult or a malfunction of the ACI algorithm.

Malfunctional ACI agents may either learn wrong policies or preserve policies in a wrong way, just like an organism that has received a high dose of ionizing radiation can no longer preserve its genetic information in a right way. Such agents can be recognized by its wrong behaviors and should be turned off and remade in most situations, just like many harmful genetic mutations drop out of the gene pool. The interval between two supervised stages should not be too long, in order to ensure the agent can still be turned off safely. However, further investigation of the wireheading problem requires a non-dualistic *gACI* model.

5. Conclusions

According to the ACI model, intelligent agents are neither optimizing machines nor ultimate wish-granters, but memories of our universe, which store the actions that have been attempted by the agents themself and their antecedents. Artificial intelligence would explore a much wider range of action space, based on the territory that has already been investigated by the lineage of *Homo sapien*.

There is growing concern that superintelligent systems created by humans may take control over the world, even drive humans into extinction. Stuart Russell (2019) argues this is like " the ancestors of the modern gorilla created (accidentally, to be sure) the genetic lineage leading to modern humans", while humans drove gorillas into an obviously worse situation, "How do the gorillas feel about this?", and calls this the *gorilla problem*.

But gorillas are not our creators. Some common ancestors about 10 million years ago created both modern gorillas and modern humans, and there's no evidence how these ancestors feel about us. Similarly, both AGI and future humans are offsprings of modern humans, genetically and/or culturally. They inherit our memories, values, and cultural traditions. If there will be a day we become a multiplanetary civilization, various approaches to coexist among humans, artificial intelligence, and cyborgs must be developed. We are not able to anticipate which approaches would be more successful than others, but we assume that all the approaches will derive from values of our human civilization, good or evil, like the mind of adults derives from their childhood.

Acknowledgements

Special thanks to Elliot Yu and Bethany Beda for their valuable help.

References

Armstrong, S. and Mindermann, S., 2018. Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in neural information processing systems*, *31*.

Bensinger, R., 2014. Solomonoff Cartesianism. *Less Wrong (blog), March*, *2*. https://www.lesswrong.com/posts/AszKwKyhBPZAnCstA/solomonoff-cartesianism

Bostrom, N., 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2), pp.71-85.

Christiano, P., 2015. Against Mimicry. *AI Alignment(blog), September, 19.* https://ai-alignment.com/against-mimicry-6002a472fc42

Demski, A. and Garrabrant, S., 2019. Embedded agency. *arXiv preprint arXiv:1902.09469*.

Dewey, D., 2011, August. Learning what to value. In *International Conference on Artificial General Intelligence* (pp. 309-314). Springer, Berlin, Heidelberg.

Good, I.J., 1966. Speculations concerning the first ultraintelligent machine. In *Advances in computers* (Vol. 6, pp. 31-88). Elsevier.

Hofstadter, D.R. and Mitchell, M., 1995. The copycat project: A model of mental fluidity and analogy-making. *Advances in connectionist and neural computation theory*, *2*, pp.205-267.

Hussein, A., Gaber, M.M., Elyan, E. and Jayne, C., 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, *50*(2), pp.1-35.

Hutter, M., 2003. A gentle introduction to the universal algorithmic agent AIXI. *Artificial General Intelligence*.

Hutter, M., Legg, S. and Vitanyi, P.M., 2007. Algorithmic probability. *Scholarpedia*, *2*(8), p.2572. http://scholarpedia.org/article/Algorithmic_probability

Kolodner, J.L., 1992. An introduction to case-based reasoning. *Artificial intelligence review*, *6*(1), pp.3-34.

Legg, S. and Hutter, M., 2007. Universal intelligence: A definition of machine intelligence. *Minds and machines*, *17*(4), pp.391-444.

Russell, S., 1998, July. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory* (pp. 101-103).

Russell, S., 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Russell, S. and Norvig, P., 2020. Artificial intelligence: a modern approach, Hoboken.

Shah, R., 2019. What is narrow value learning?. *Less Wrong (blog)*, *January*, *10*. https://www.lesswrong.com/posts/vX7KirQwHsBaSEdfK/what-is-narrow-value-learning

Solomonoff, R.J., 1964. A formal theory of inductive inference. Part II. *Information and control*, 7(2), pp.224-254.

Yudkowsky, E., 2011. Complex value systems are required to realize valuable futures. The Singularity Institute, San Francisco, CA.