
S FRANCISCO
D'ASSIS

Naturalized Induction

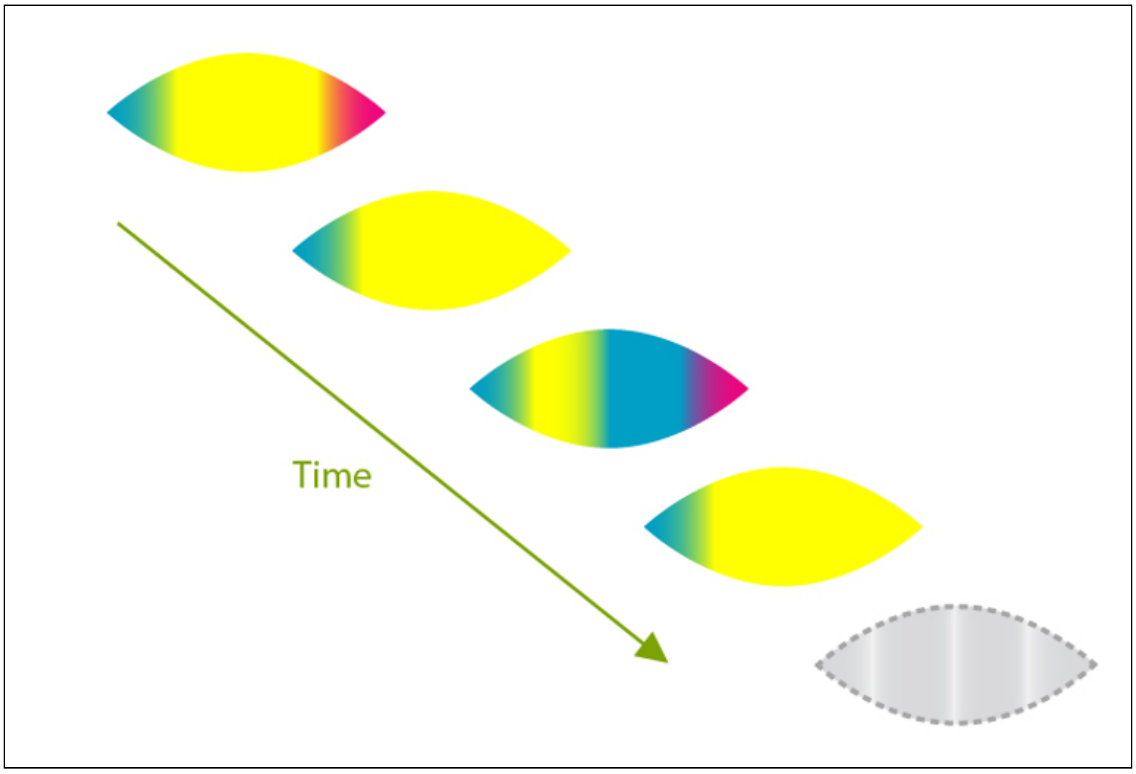
Building Phenomenological Bridges

Naturalized induction is an open problem in Friendly Artificial Intelligence (OPFAI). The problem, in brief: Our current leading models of induction do not allow reasoners to treat their own computations as processes in the world.

The problem's roots lie in algorithmic information theory and formal epistemology, but finding answers will require us to wade into debates on everything from theoretical physics to anthropic reasoning and self-reference. This post will lay the groundwork for a sequence of posts (titled '**Artificial Naturalism**') introducing different aspects of this OPFAI.

AI perception and belief: A toy model

A more concrete problem: Construct an algorithm that, given a sequence of the colors cyan, magenta, and yellow, predicts the next colored field.



Colors: *CYYM* *CYYY* *CYCM* *CYYY* *????*

This is an instance of the general problem 'From an incomplete data series, how can a reasoner best make predictions about future data?'. In practice, any agent that acquires

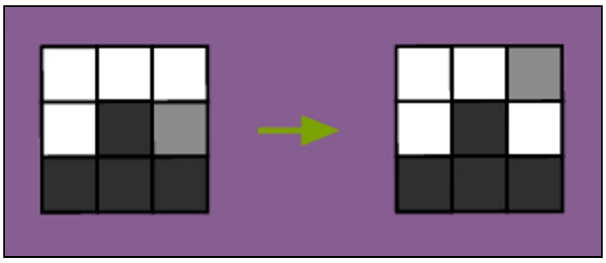
information from its environment and makes predictions about what's coming next will need to have two map-like[1] subprocesses:

1. Something that generates the agent's predictions, its expectations. By analogy with human scientists, we can call this prediction-generator the agent's **hypotheses** or **beliefs**.

2. Something that transmits new information to the agent's prediction-generator so that its hypotheses can be updated. Employing another anthropomorphic analogy, we can call this process the agent's **data** or **perceptions**.

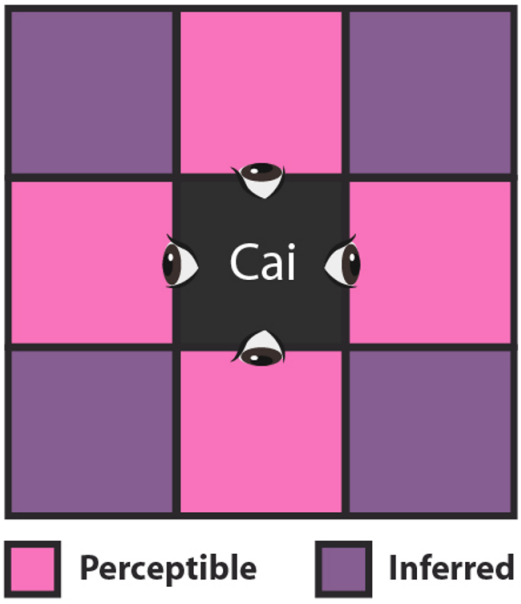
Here's an example of a hypothesis an agent could use to try to predict the next color field. I'll call the imaginary agent '**Cai**'. Any reasoner will need to begin with some (perhaps provisional) assumptions about the world.[2] Cai begins with the belief[3] that its environment behaves like a cellular automaton: the world is a grid whose tiles change over time based on a set of stable laws. The laws are local in time and space, meaning that you can perfectly predict a tile's state based on the states of the tiles next to it a moment prior — if you know which laws are in force.

Cai believes that it lives in a closed 3x3 grid where tiles have no diagonal effects. Each tile can occupy one of three states. We might call the states '0', '1', and '2', or, to make visualization easier, 'white', 'black', and 'gray'. So, on Cai's view, the world as it changes looks something like this:



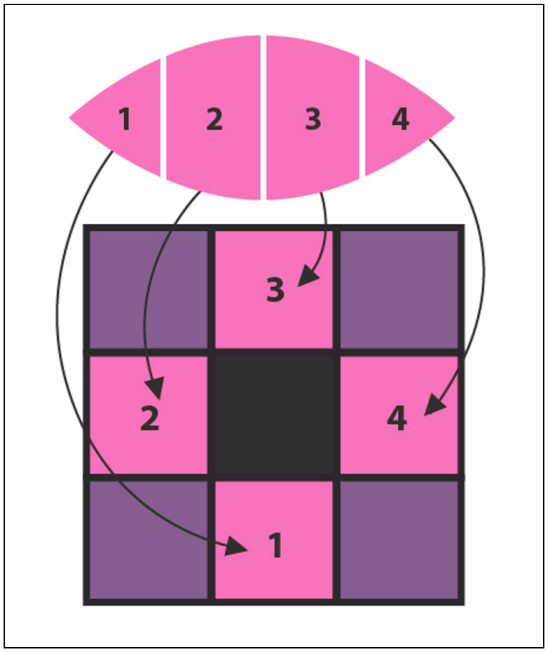
An example of the world's state at one moment, and its state a moment later.

Cai also has beliefs about its own location in the cellular automaton. Cai believes that it is a black tile at the center of the grid. Since there are no diagonal laws of physics in this world, Cai can only directly interact with the four tiles directly above, below, to the left, and to the right. As such, any perceptual data Cai acquires will need to come from those four tiles; anything else about Cai's universe will be known only by inference.



Cai perceives stimuli in four directions. Unobservable tiles fall outside the cross.

How does all this bear on the color-predicting problem? Cai hypothesizes that the sequence of colors is sensory — it's an experience within Cai, triggered by environmental changes. Cai conjectures that since its visual field comes in at most four colors, its visual field's quadrants probably represent its four adjacent tiles. The leftmost color comes from a southern stimulus, the next one to the right from a western stimulus, then a northern one, then an eastern one. And the south, west, north, east cycle repeats again and again.

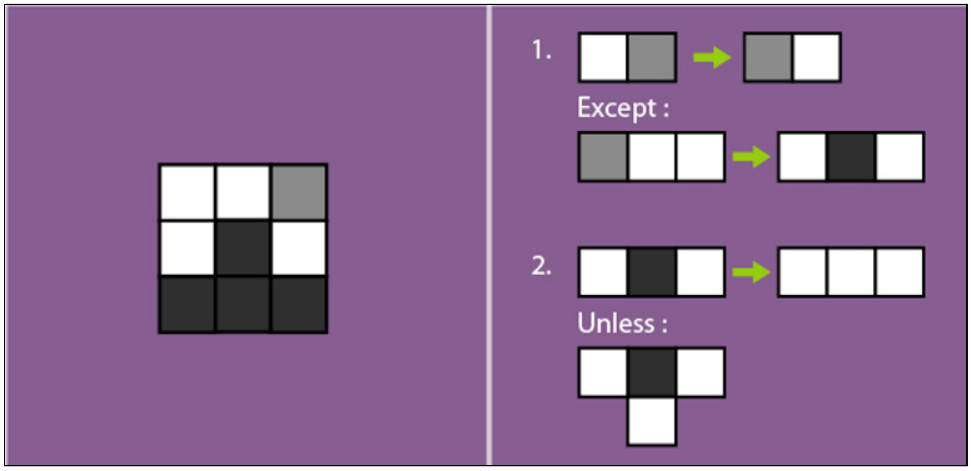


Cai’s visual experiences break down into quadrants, corresponding to four directions.

On this model, the way Cai’s senses organize the data isn't wholly veridical; the four patches of color aren’t perfectly shaped like Cai’s environment. But the organization of Cai's sensory apparatus and the organization of the world around Cai are similar enough that Cai can reconstruct many features of its world.

By linking its visual patterns to patterns of changing tiles, Cai can hypothesize laws that guide the world's changes and explain Cai's sensory experiences. Here's one possibility, Hypothesis A:

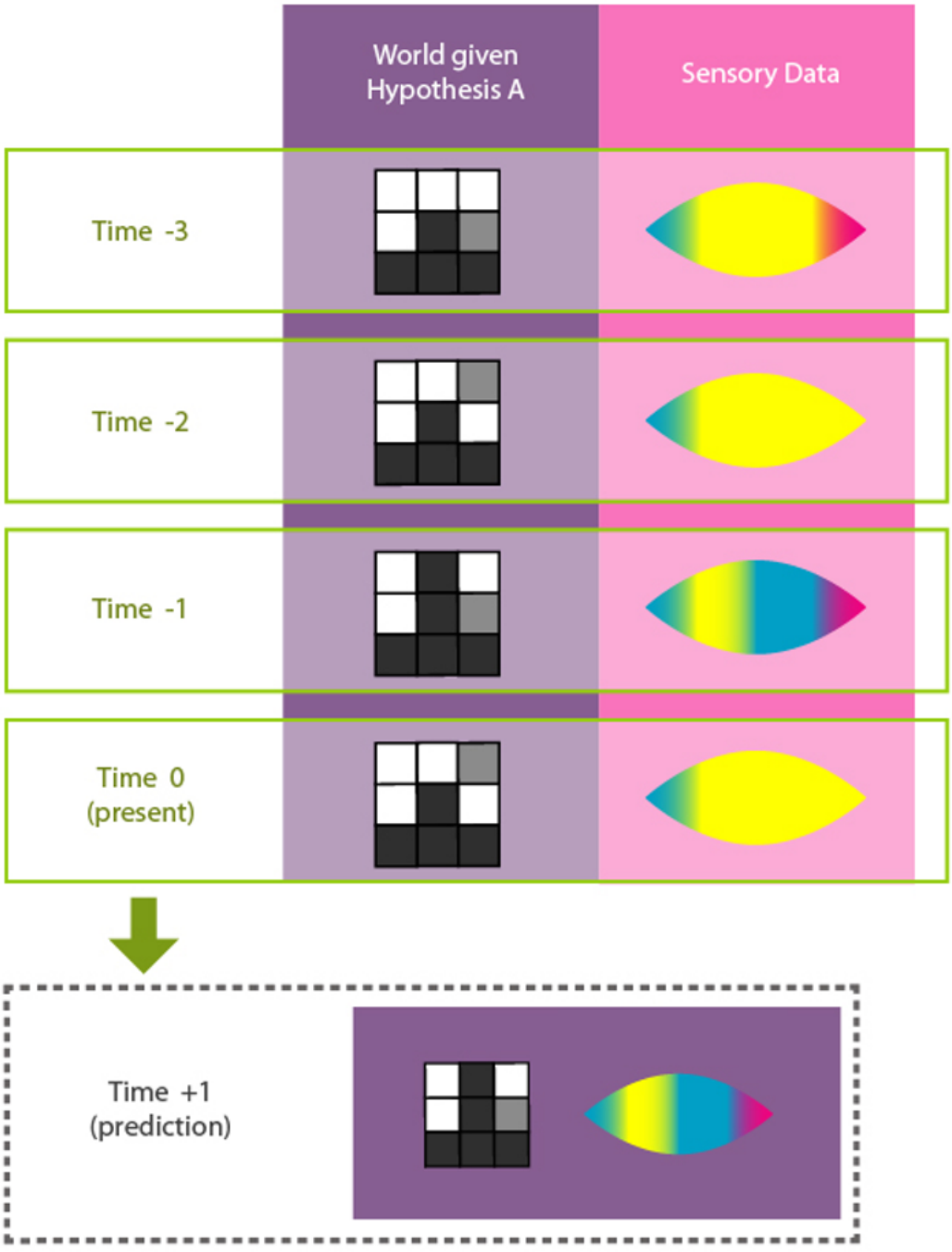
- Black corresponds to cyan, white to yellow, and gray to magenta.
- At present, the top two rows are white and the bottom row is black, except for the upper-right tile (which is gray) and Cai itself, a black middle tile.
- Adjacent gray and white tiles exchange shades. Exception: When a white tile is pinned by a white and gray tile on either side, it turns black.
- Black tiles pinned by white ones on either side turn white. Exception: When the black tile is adjacent to a third white tile, it remains black.



Hypothesis A's physical content. On the left: Cai's belief about the world's present state. On the right: Cai's belief about the rules by which the world changes over time. The rules are symmetric under rotation and reflection.

Bridging stimulus and experience

So that's one way of modeling Cai's world; and it will yield a prediction about the cellular automaton's next state, and therefore about Cai's next visual experience. It will also yield retrodictions of the cellular automaton's state during Cai's three past sensory experiences.

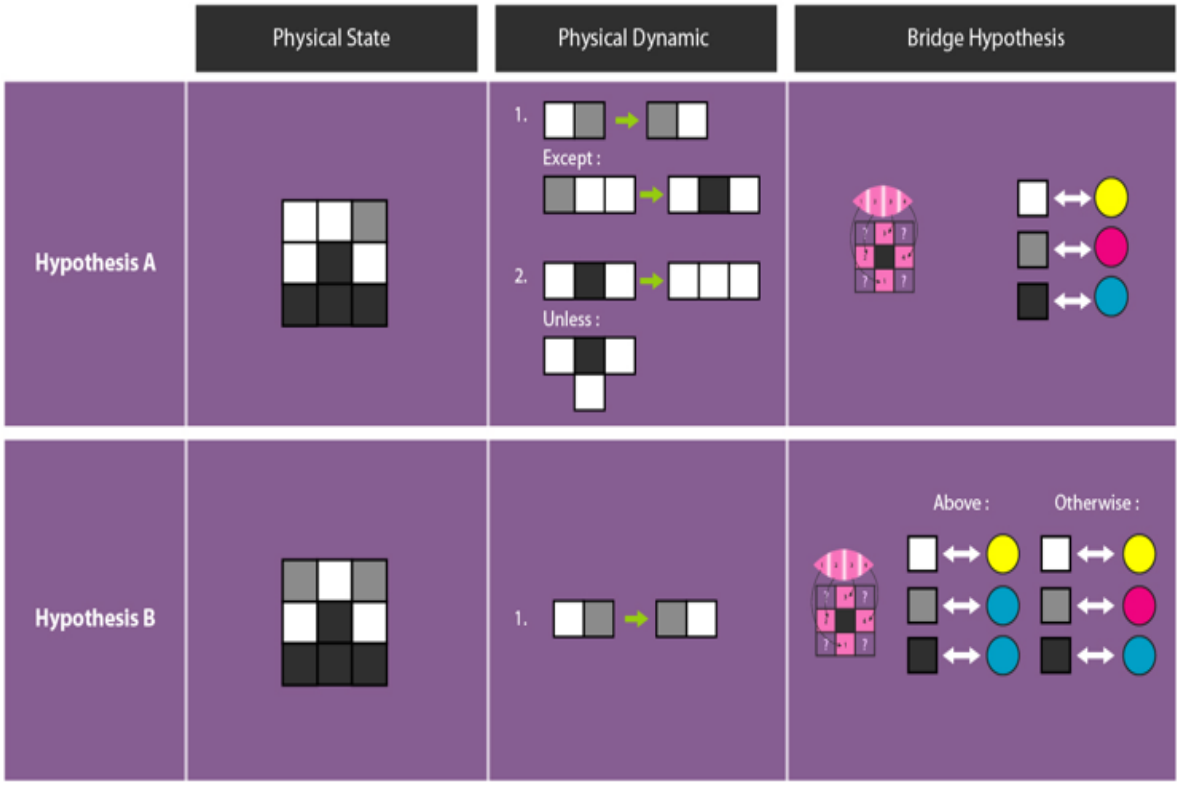


Hypothesis A asserts that tiles below Cai, to Cai's left, above, and to Cai's right relate to Cai's color experiences via the rule {black ↔ cyan, white ↔ yellow, gray ↔ magenta}. Corner tiles, and future world-states and experiences, can be inferred from Hypothesis A's cell transition rules.

Are there other, similar hypotheses that can explain the same data? Here's one, Hypothesis B:

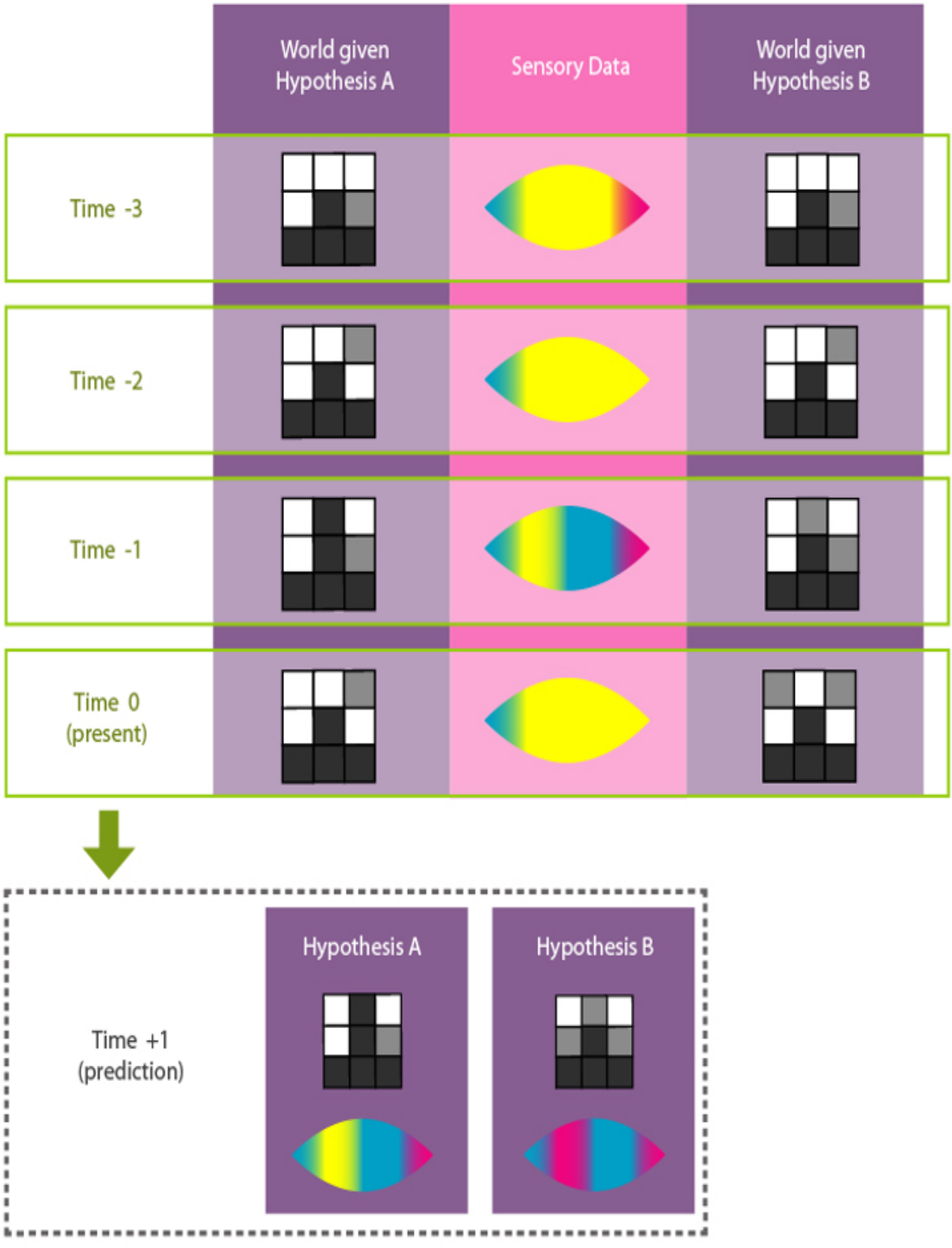
- Normally, the correspondences between experienced colors and neighboring tile states are {black ↔ cyan, white ↔ yellow, gray ↔ magenta}, as in Hypothesis A. But northern grays are perceived as though they were black, helping explain irregularities in the distribution of cyan.
- Hypothesis B's cellular automaton presently looks similar to Hypothesis A's, but with a gray tile in the upper-left corner.
- Adjacent gray and white tiles exchange shades. Nothing else changes.

The added complexity in the perception-to-environment link allows Hypothesis B to do away with most of the complexity in Hypothesis A's physical laws. Breaking down Hypotheses A and B into their respective physical and perception-to-environment components makes it more obvious how the two differ:



A has the simpler bridge hypothesis, while B has the simpler physical hypothesis.

Though they share a lot in common, and both account for Cai's experiences to date, these two hypotheses diverge substantially in the cellular automaton states and future experiences they predict:



The two hypotheses infer different distributions and dynamical rules for the tile shades from the same perceptual data. These worldly differences then diverge in the future experiences they predict.

Hypotheses linking observations to theorized entities appear to be quite different from hypothesis that just describe the theorized entities in their own right. In Cai's case, the latter hypotheses look like pictures of physical worlds, while the former are ties between different kinds of representation. But in both cases it's useful to treat these processes in humans or machines as beliefs, since they can be assigned weights of expectation and updated.

'Phenomenology' is a general term for an agent's models of its own introspected experiences. As such, we can call these hypotheses linking experienced data to theorized processes **phenomenological bridge hypotheses**. Or just 'bridge hypotheses', for short.

If we want to build an agent that tries to evaluate the accuracy of a model based on the accuracy of its predictions, we need some scheme to compare thingies in the model (like tiles) and thingies in the sensory stream (like colors). Thus a **bridge rule** appears to be necessary to talk about induction over models of the world. And bridge hypotheses are just bridge rules treated as probabilistic, updatable beliefs.

As the last figure above illustrates, bridge hypotheses can make a big difference for one's scientific beliefs and expectations. And bridge hypotheses aren't a free lunch; it would be a mistake to shunt all complexity onto them in order to simplify your physical hypotheses. Allow your bridge hypotheses to get too complicated, and you'll be able to justify mad world-models, e.g., ones where the universe consists of a single apricot whose individual atoms each get a separate bridge to some complex experience. At the same time, if you demand *too much* simplicity from your bridge hypotheses, you'll end up concluding that the physical world consists of a series of objects shaped just like your mental states. That way you can get away with a comically simple bridge rule like {exists(x) ↔ experiences(y,x)}.

In the absence of further information, it may not be possible to rule out Hypothesis A or Hypothesis B. The takeaway is that tradeoffs between the complexity of bridging hypotheses and the complexity of physical hypotheses do occur, and do matter. Any artificial agent needs some way of formulating good hypotheses of this type in order to be able to understand the universe at all, whether or not it finds itself in doubt after it has done so.

Generalizing bridge rules and data

Reasoners — both human and artificial — don't begin with perfect knowledge of their own design. When they have working self-models at all, these self-models are fallible. Aristotle thought the brain was an organ for cooling the blood. We had to find out about neurons by opening up the heads of people who looked like us, putting the big corrugated gray organ under a microscope, seeing (with our eyes, our visual cortex, our senses) that the microscope (which we'd previously generalized shows us tiny things as if they were large) showed this incredibly fine mesh of connected blobs, and realizing, "Hey, I bet this does information processing and *that's* what I am! The big gray corrugated organ that's inside my own head is *me*!"

The bridge hypotheses in Hypotheses A and B are about linking an agent's environment-triggered experiences to environmental causes. But in fact bridge hypotheses are more general than that.

1. An agent's experiences needn't all have *environmental* causes. They can be caused by something inside the agent.

2. The cause-effect relation we're bridging can go the other way. E.g., a bridge hypothesis can link an experienced *decision* to a behavioral consequence, or to an expected outcome of the behavior.

3. The bridge hypothesis needn't link causes to effects at all. E.g., it can assert that the agent's experienced sensations or decisions just *are* a certain physical state. Or it can assert neutral correlations.

Phenomenological bridge hypotheses, then, can relate theoretical posits to any sort of experiential data. Experiential data are internally evident facts that get compared to hypotheses and cause updates — the kind of data of direct epistemic relevance to individual scientists updating their personal beliefs. Light shines on your retina, gets transduced to neural firings, gets reconstructed in your visual cortex and then — this is the key part — that internal fact gets used to decide what sort of universe you're probably in.

The data from an AI's environment is just one of many kinds of information it can use to update its probability distributions. In addition to ordinary sensory content such as vision and smell, update-triggering data could include things like how much RAM is being used. This is because an inner RAM sense can tell you that the universe is such as to include a copy of you with at least that much RAM.

We normally think of science as reliant mainly on sensory faculties, not introspective ones. Arriving at conclusions just by examining your own intuitions and imaginings sounds more like math or philosophy. But for present purposes the distinction isn't important. What matters is just whether the AGI forms accurate beliefs and makes good decisions. Prototypical scientists may shun introspectionism because humans do a better job of directly apprehending and communicating facts about their environments than facts about their own inner lives, but AGIs can have a very different set of strengths and weaknesses. Although introspection, like sensation, is fallible, introspective self-representations sometimes empirically correlate with world-states.[4] And that's all it takes for them to constitute Bayesian evidence.

Bridging hardware and experience

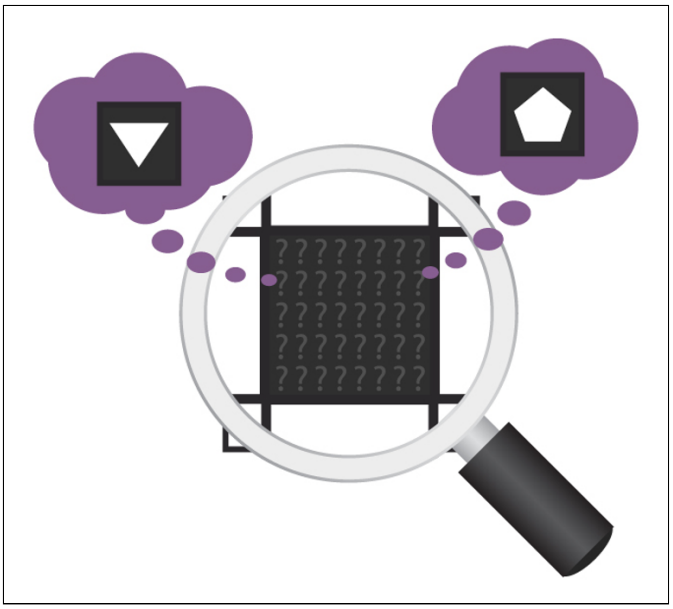
In my above discussion, all of Cai's world-models included representations of Cai itself. However, these representations were very simple — no more than a black tile in a specific environment. Since Cai's own computations are complex, it must be the case that either they are occurring outside the universe depicted (as though Cai is plugged into a cellular automaton Matrix), or the universe depicted is much more complex than Cai thinks.[5] Perhaps its model is wildly mistaken, or perhaps the high-level cellular patterns it's hypothesized arise from other, smaller-scale regularities.

Regardless, Cai's computations must be embodied in some causal pattern. Cai will eventually need to construct bridge hypotheses between its experiences and their physical substrate if it is to make reliable predictions about its own behavior and about its relationship with its surroundings.

Visualize the epistemic problem that an agent needs to solve. Cai has access to a series of sensory impressions. In principle we could also add introspective data to that. But

you'll still get a series of (presumably time-indexed) facts in some native format of that mind. Those facts very likely won't be structured exactly like any ontologically basic feature of the universe in which the mind lives. They won't be a precise position of a Newtonian particle, for example. And even if we were dealing with sense data shaped just like ontologically basic facts, a rational agent could never know for certain that they were ontologically basic, so it would still have to consider hypotheses about even more basic particles.

When humans or AGIs try to match up hypotheses about universes to sensory experiences, there will be a type error. Our representation of the universe will be in hypothetical atoms or quantum fields, while our representation of sensory experiences will be in a native format like 'red-green'.[6] This is where bridge rules like Cai's color conversions come in — bridges that relate our experiences to environmental stimuli, as well as ones that relate our experiences to the hardware that runs us.



Cai can form physical hypotheses about its own internal state, in addition to ones about its environment. This means it can form bridge hypotheses between its experiences and its own hardware, in addition to ones between its experiences and environment.

If you were an AI, you might be able to decode your red-green visual field into binary data — on-vs.-off — and make very simple hypotheses about how that corresponded to transistors making you up. Once you used a microscope on yourself to see the transistors, you'd see that they had binary states of positive and negative voltage, and all that would be left would be a hypothesis about whether the positive (or negative) voltage corresponded to an introspected 1 (or 0).

But even then, I don't quite see how you could do without the bridge rules — there has to be some way to go from internal sensory types to the types featured in your hypotheses about physical laws.

Our sensory experience of red, green, blue *is* certain neurons firing in the visual cortex, and these neurons are in turn made from atoms. But internally, so far as information processing goes, we just know about the red, the green, the blue. This is what you'd expect an agent made of atoms to feel like from the inside. Our native representation of a pixel field won't come with a little tag telling us with infallible transparency about the underlying quantum mechanics.

But this means that when we're done positing a physical universe in all its detail, we also need one last (hopefully simple!) step that connects hypotheses about 'a brain that processes visual information' to 'I see blue'.

One way to avoid worrying about bridge hypotheses would be to instead code the AI to accept **bridge axioms**, bridge rules with no degrees of freedom and no uncertainty. But the AI's designers are not in fact infinitely confident about how the AI's perceptual states emerge from the physical world — that, say, quantum field theory is the One True Answer, and shall be so from now until the end of time. Nor can they transmit infinite rational confidence to the AI merely by making it more stubbornly convinced of the view. If you pretend to know more than you do, the world will still bite back. As an agent in the world, you really do have to think about and test a variety of different uncertain hypotheses about what hardware you're running on, what kinds of environmental triggers produce such-and-such experiences, and so on. This is particularly true if your hardware is likely to undergo substantial changes over time.

If you don't allow the AI to form probabilistic, updatable hypotheses about the relation between its phenomenology and the physical world, the AI will either be unable to reason at all, or it will reason its way off a cliff. In my next post, **Bridge Collapse**, I'll begin discussing how the latter problem sinks an otherwise extremely promising approach to formalizing ideal AGI reasoning: Solomonoff induction.

---

[1] By 'map-like', I mean that the processes look similar to the representational processes in human thought. They systematically correlate with external events, within a pattern-tracking system that can readily propagate and exploit the correlation. ↩

[2] Agents need initial assumptions, built-in prior information. The prior is defined by whatever algorithm the reasoner follows in making its very first updates.

If I leave an agent's priors undefined, no ghost of reasonableness will intervene to give the agent a 'default' prior. For example, it won't default to a uniform prior over possible coinflip outcomes in the absence of relevant evidence. Rather, without something that acts like a prior, the agent just won't work — in the same way that a calculator won't work if you grant it the freedom to do math however it wishes. A frequentist AI might refuse to *talk* about priors, but it would still need to *act* like it has priors, else break. ↩

[3] This talk of 'belief' and 'assumption' and 'perception' *is* anthropomorphizing, and the analogies to human psychology won't be perfect. This is important to keep in view, though there's only so much we can do to avoid vagueness and analogical reasoning when the architecture of AGIs remains unknown. In particular, I'm not assuming that every artificial scientist is particularly intelligent. Or particularly conscious.

What I mean with all this 'Cai believes...' talk is that Cai weights predictions and selects actions *just as though* it believed itself to be in a cellular automaton world. One can treat Cai's automaton-theoretic model as just a bookkeeping device for assigning Cox's-theorem-following real numbers to encoded images of color fields. But one can also treat Cai's model as a psychological expectation, to the extent it functionally resembles the corresponding human mental states. Words like 'assumption' and 'thinks' here needn't mean that the agent thinks in the same fashion humans think; what we're interested in are the broad class of information-processing algorithms that yield similar behaviors. ↩

4 To illustrate: In principle, even a human pining to become a parent could, by introspection alone, infer that they might be an evolved mind (since they are experiencing a desire to self-replicate) and embedded in a universe which had evolved minds with evolutionary histories. An AGI with more reliable internal monitors could learn a great deal about the rest of the universe just by investigating itself. ↩

5 In either case, we shouldn't be surprised to see Cai failing to fully represent its own inner workings. An agent cannot explicitly represent itself in its totality, since it would then need to represent itself representing itself representing itself ... ad infinitum. Environmental phenomena, too, must usually be compressed. ↩

6 One response would be to place the blame on Cai's positing white, gray, and black for its world-models, rather than sticking with cyan, yellow, and magenta. But there will still be a type error when one tries to compare perceived cyan/yellow/magenta with hypothesized (but perceptually invisible) cyan/yellow/magenta. Explicitly introducing separate words for hypothesized v. perceived colors doesn't produce the distinction; it just makes it easier to keep track of a distinction that was already present. ↩

Bridge Collapse: Reductionism as Engineering Problem

Followup to: Building Phenomenological Bridges

Summary: AI theorists often use models in which agents are crisply separated from their environments. This simplifying assumption can be useful, but it leads to trouble when we build machines that presuppose it. A machine that believes it can only interact with its environment in a narrow, fixed set of ways will not understand the value, or the dangers, of self-modification. By analogy with Descartes' mind/body dualism, I refer to agent/environment dualism as *Cartesianism*. The open problem in Friendly AI (OPFAI) I'm calling naturalized induction is the project of replacing Cartesian approaches to scientific induction with reductive, physicalistic ones.

I'll begin with a story about a storyteller.

Once upon a time — specifically, 1976 — there was an AI named TALE-SPIN. This AI told stories by inferring how characters would respond to problems from background knowledge about the characters' traits. One day, TALE-SPIN constructed a most peculiar tale.

> Henry Ant was thirsty. He walked over to the river bank where his good friend Bill Bird was sitting. Henry slipped and fell in the river. Gravity drowned.

Since Henry fell in the river near his friend Bill, TALE-SPIN concluded that Bill rescued Henry. But for Henry to fall in the river, gravity must have pulled Henry. Which means gravity must have been in the river. TALE-SPIN had never been told that gravity knows how to swim; and TALE-SPIN had never been told that gravity has any friends. So gravity drowned.

TALE-SPIN had previously been programmed to understand involuntary motion in the case of characters being pulled or carried by other characters — like Bill rescuing Henry. So it was programmed to understand 'character X fell to place Y' as 'gravity moves X to Y', as though gravity were a character in the story.[1]

For us, the hypothesis 'gravity drowned' has low prior probability because we know gravity isn't the *type* of thing that swims or breathes or makes friends. We want agents to seriously consider whether the law of gravity pulls down rocks; we don't want agents to seriously consider whether the law of gravity pulls down the law of electromagnetism. We may not want an AI to assign *zero* probability to 'gravity drowned', but we at least want it to neglect the possibility as Ridiculous-By-Default.

When we introduce deep type distinctions, however, we also introduce new ways our stories can fail.

Hutter's cybernetic agent model

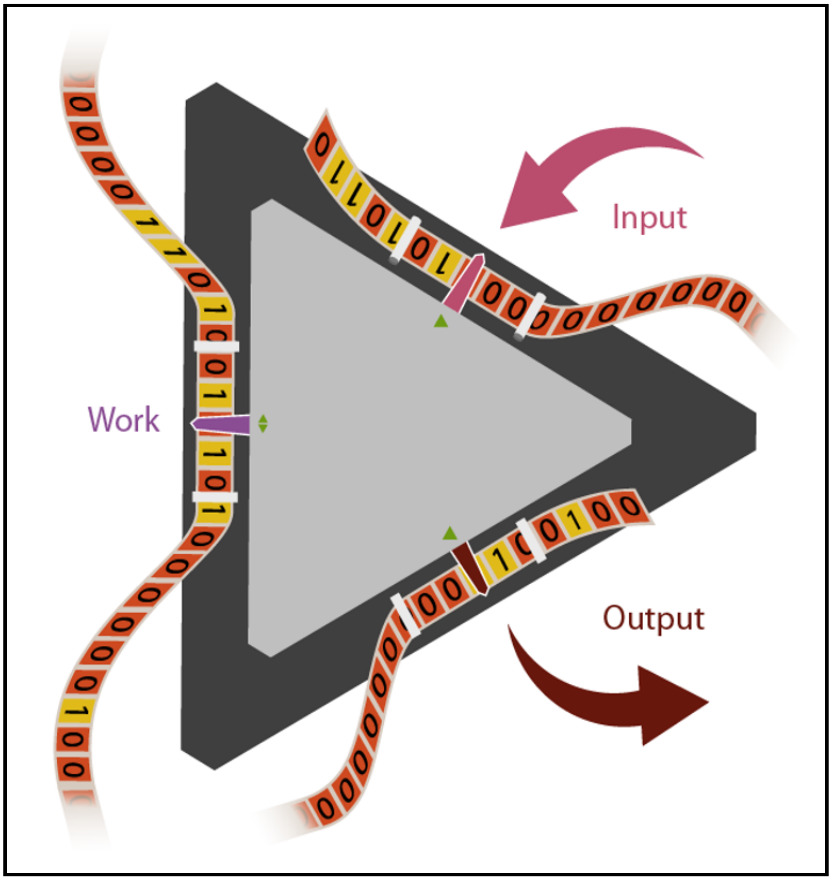
Russell and Norvig's leading AI textbook credits Solomonoff with setting the agenda for the field of AGI: "AGI looks for a universal algorithm for learning and acting in any environment, and has its roots in the work of Ray Solomonoff[.]" As an approach to AGI, Solomonoff induction presupposes a model with a strong type distinction between the 'agent' and the 'environment'. To make its intuitive appeal and attendant problems more obvious, I'll sketch out the model.

A Solomonoff-inspired AI can most easily be represented as a multi-tape Turing machine like the one Alex Altair describes in An Intuitive Explanation of Solomonoff Induction. The machine has:

- three tapes, labeled 'input', 'work', and 'output'. Each initially has an infinite strip of 0s written in discrete cells.

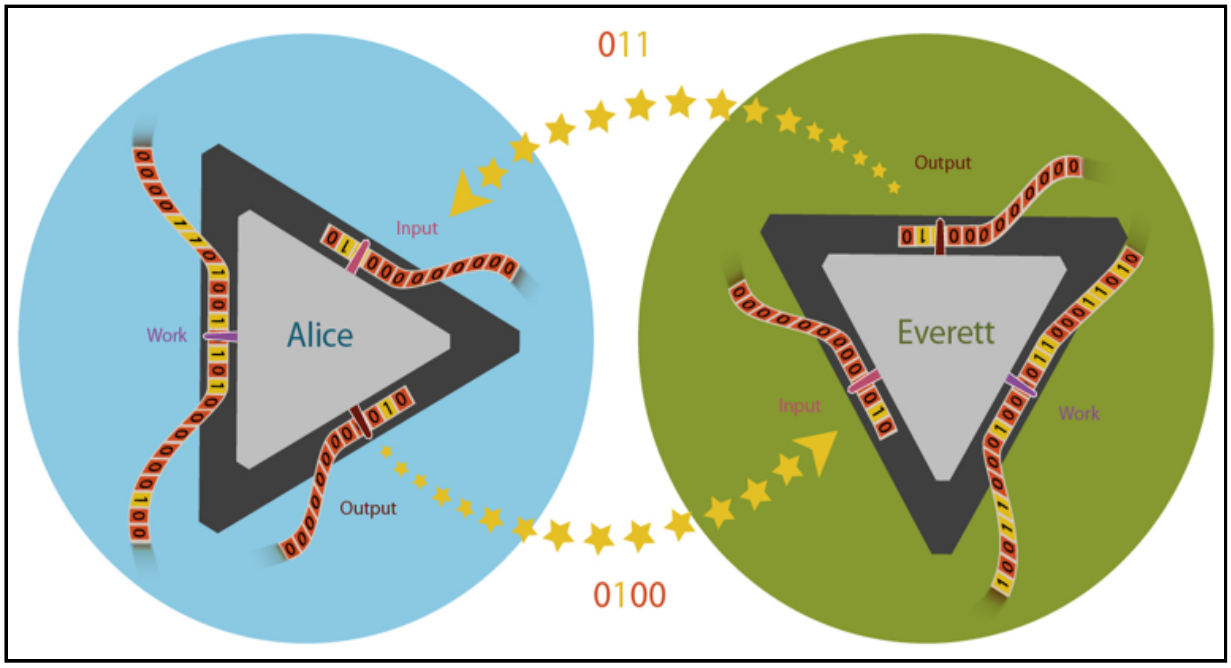
- one head per tape, with the input head able to read its cell's digit and move to the right, the output head able to write 0 or 1 to its cell and move to the right, and the work head able to read, write, move, or do nothing, and how to transition to another rule.

- a program, consisting of a finite, fixed set of transition rules. Each rule says when heads read, write, move, or do nothing, and how to transition to another rule.



A three-tape Turing machine.

We could imagine two such Turing machines communicating with each other. Call them 'Agent' and 'Environment', or 'Alice' and 'Everett'. Alice and Everett take turns acting. After Everett writes a bit to his output tape, that bit magically appears on Alice's input tape; and likewise, when Alice writes to her output tape, it gets copied to Everett's input tape. AI theorists have used this setup, which Marcus Hutter calls the **cybernetic agent model**, as an extremely simple representation of an agent that can **perceive** its environment (using the input tape), **think** (using the work tape), and **act** (using the output tape).[2]



A Turing machine model of agent-environment interactions. At first, the machines differ only in their programs. 'Alice' is the agent we want to build, while 'Everett' stands for everything else that's causally relevant to Alice's success.

We can define Alice and Everett's behavior in terms of any bit-producing Turing machines we'd like, including ones that represent probability distributions and do Bayesian updating. Alice might, for example, use her work tape to track four distinct possibilities and update probabilities over them:[3]

- (a) Everett always outputs 0.
- (b) Everett always outputs 1.
- (c) Everett outputs its input.
- (d) Everett outputs the opposite of its input.

Alice starts with a uniform prior, i.e., 25% probability each. If Alice's first output is 1, and Everett responds with 1, then Alice can store those two facts on her work tape and conditionalize on them both, treating them as though they were certain. This results in 0.5 probability each for (b) and (c), 0 probability for (a) and (d).

We care about an AI's epistemology only because it informs the AI's behavior — on this model, its bit output. If Alice outputs whatever bits maximize her expected chance of receiving 1s as input, then we can say that Alice **prefers** to perceive 1. In the example I just gave, such a preference predicts that Alice will proceed to output 1 forever. Further exploration is unnecessary, since she knows of no other importantly different hypotheses to test.

Enriching Alice's set of hypotheses for how Everett could act will let Alice win more games against a wider variety of Turing machines. The more programs Alice can pick out and assign a probability to, the more Turing machines Alice will be able to identify and intelligently respond to. If we aren't worried about whether it takes Alice ten minutes or a billion years to compute an update, and Everett will always patiently wait his turn, then we can simply have Alice perform perfect Bayesian updates; if her priors are right, and she translates her beliefs into sensible actions, she'll then be able to optimally respond to any environmental Turing machine.

For AI researchers following Solomonoff's lead, that's the name of the game: Figure out the program that will let Alice behave optimally while communicating with as wide a range of Turing machines as possible, and you've at least solved the *theoretical* problem of picking out the optimal artificial agent from the space of possible reasoners. The agent/environment model here may look simple, but a number of theorists see it as distilling into its most basic form the task of an AGI.[2]

Yet a Turing machine, like a cellular automaton, is an abstract machine — a creature of thought experiments and mathematical proofs. Physical computers can act like abstract computers, in just the same sense that heaps of apples can behave like the abstract objects we call 'numbers'. But computers and apples are high-level generalizations, imperfectly represented by concise equations.[4] When we move from our mental models to trying to build an actual AI, we have to pause and ask how well our formalism captures what's going on in reality.

The problem with Alice

'Sensory input' or 'data' is what I call the information Alice conditionalizes on; and 'beliefs' or 'hypotheses' is what I call the resultant probability distribution and representation of possibilities (in Alice's program or work tape). This distinction seems basic to reasoning, so I endorse programming agents to treat them as two clearly distinct types. But in building such agents, we introduce the possibility of **Cartesianism**.

René Descartes held that human minds and brains, although able to causally interact with each other, can each exist in the absence of the other; and, moreover, that the properties of purely material things can never fully explain minds. In his honor, we can call a model or procedure *Cartesian* if it treats the reasoner as a being separated from the physical universe. Such a being can perceive (and perhaps alter) physical processes, but it can't be identified with any such process.[5]

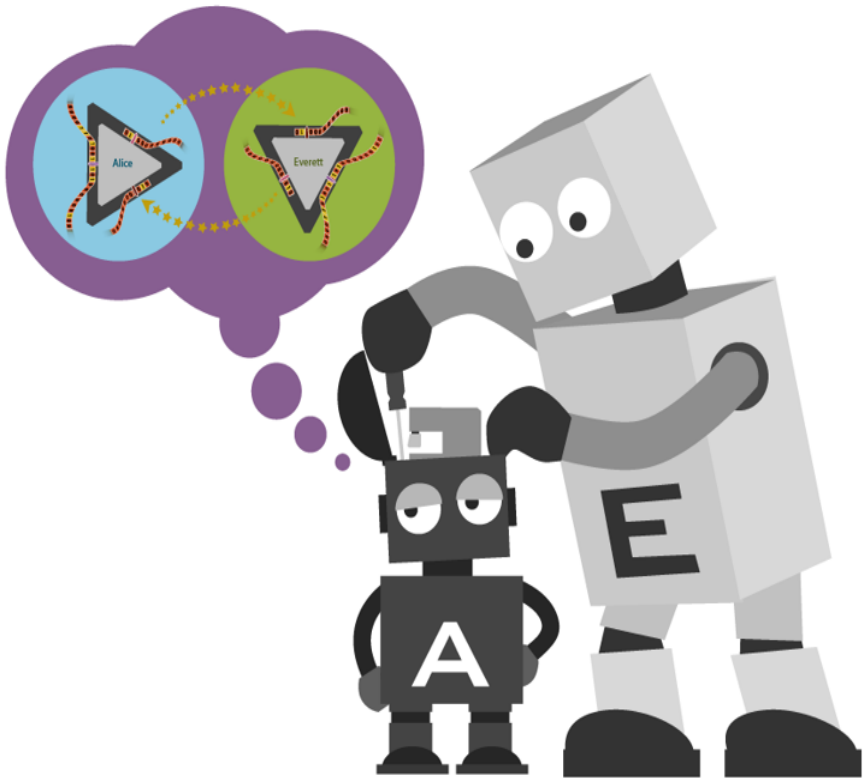
The relevance of Cartesians to AGI work is that we can model them as agents experiencing a strong type distinction between 'mind' and 'matter', and an unshakable belief in the metaphysical independence of those two categories; *because* they're of such different kinds, they can vary independently. So we end up with AI errors that are the opposite of TALE-SPIN's — like an induction procedure that distinguishes gravity's type from embodied characters' types *so* strongly that it cannot hypothesize that, say, particles underlie or mediate both phenomena.

My claim is that if we plug in 'Alice's sensory data' for 'mind' and 'the stuff Alice hypothesizes as causing the sensory data' for 'matter', then agents that can only model themselves using the cybernetic agent model are Cartesian in the relevant sense.[6]

The model is Cartesian because the agent and its environment *can only interact by communicating*. That is, their only way of affecting each other is by trading bits printed to tapes.

If we build an actual AI that believes it's like Alice, it will believe that the environment can't affect it in ways that aren't immediately detectable, can't edit its source code, and can't force it to halt. But that makes the Alice-Everett system almost nothing like a physical agent embedded in a real environment. Under many circumstances, a real AI's environment will alter it directly. E.g., the AI can fall into a volcano. A volcano doesn't harm the agent by feeding unhelpful bits into its environmental sensors. It harms the agent by destroying it.

A more naturalistic model would say: Alice outputs a bit; Everett reads it; and then Everett does whatever the heck he wants. That might be feeding a new bit into Alice. Or it might be vandalizing Alice's work tape, or smashing Alice flat.



A robotic Everett tampering with an agent that mistakenly assumes Cartesianism. A real-world agent's computational states have physical correlates that can be directly edited by the environment. If the agent can't model such scenarios, its reasoning (and resultant decision-making) will suffer.

A still more naturalistic approach would be to place Alice *inside* of Everett, as a subsystem. In the real world, agents are surrounded by their environments. The two form a cohesive whole, bound by the same physical laws, freely interacting and commingling.

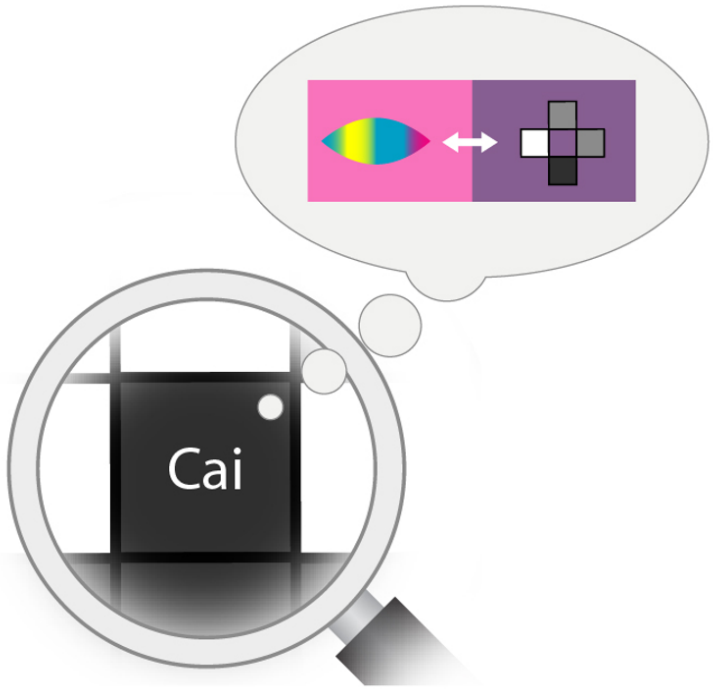
If Alice only worries about whether Everett will output a 0 or 1 to her sensory tape, then no matter how complex an understanding Alice has of Everett's inner workings, Alice will fundamentally misunderstand the situation she's in. Alice won't be able to represent hypotheses about how, for example, a pill might erase her memories or otherwise modify her source code.

Humans, in contrast, can readily imagine a pill that modifies our memories. It seems childishly easy to hypothesize being changed by avenues other than perceived sensory information. The limitations of the cybernetic agent model aren't immediately obvious, because it isn't easy for us to put ourselves in the shoes of agents with alien blind spots.

There *is* an agent-environment distinction, but it's a pragmatic and artificial one. The boundary between the part of the world we call 'agent' and the part we call 'not-agent' (= 'environment') is frequently fuzzy and mutable. If we want to build an agent that's robust across many environments and self-modifications, we can't just design a program that excels at predicting sensory sequences generated by Turing machines. We need an agent that can form accurate beliefs about the actual world it lives in, including accurate beliefs about its own physical underpinnings.

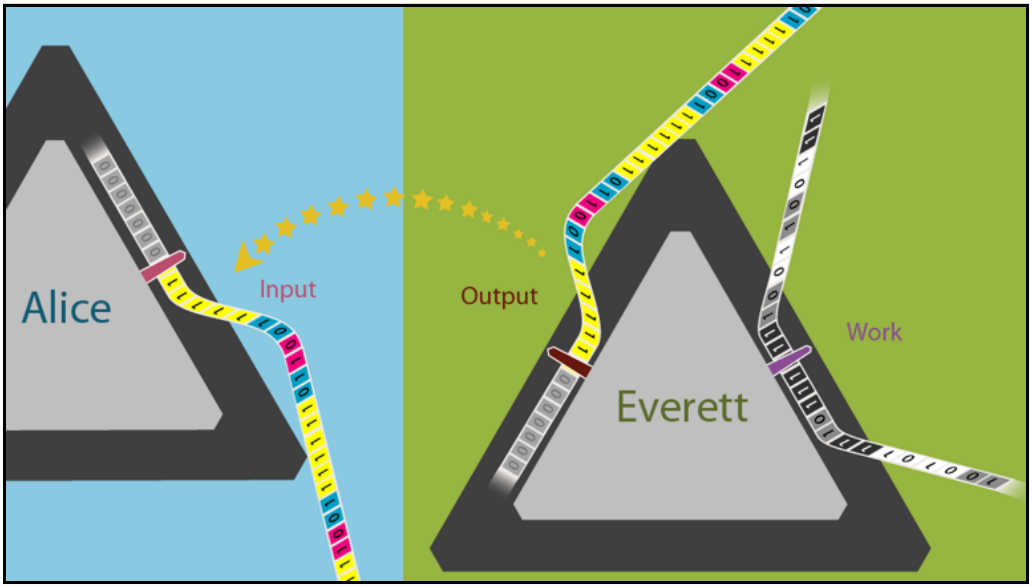
From Cartesianism to naturalism

What would a naturalized self-model, a model of the agent as a process embedded in a lawful universe, look like? As a first attempt, one might point to the pictures of Cai in Building Phenomenological Bridges.



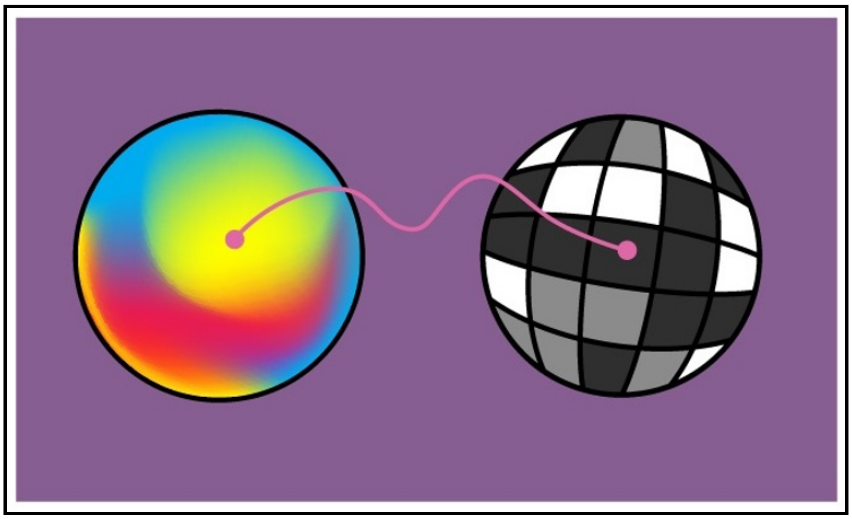
Cai has a simple physical model of itself as a black tile at the center of a cellular automaton grid. Cai's phenomenological bridge hypotheses relate its sensory data to surrounding tiles' states.

But this doesn't yet specify a non-Cartesian agent. To treat Cai as a Cartesian, we could view the tiles surrounding Cai as the work tape of Everett, and the dynamics of Cai's environment as Everett's program. (We can also convert Cai's perceptual experiences into a binary sequence on Alice/Cai's input tape, with a translation like 'cyan = 01, magenta = 10, yellow = 11'.)



Alice/Cai as a cybernetic agent in a Turing machine circuit.

The problem isn't that Cai's world is Turing-computable, of course. It's that if Cai's hypotheses are solely about what sorts of perception-correlated patterns of environmental change can occur, then Cai's models will be Cartesian.

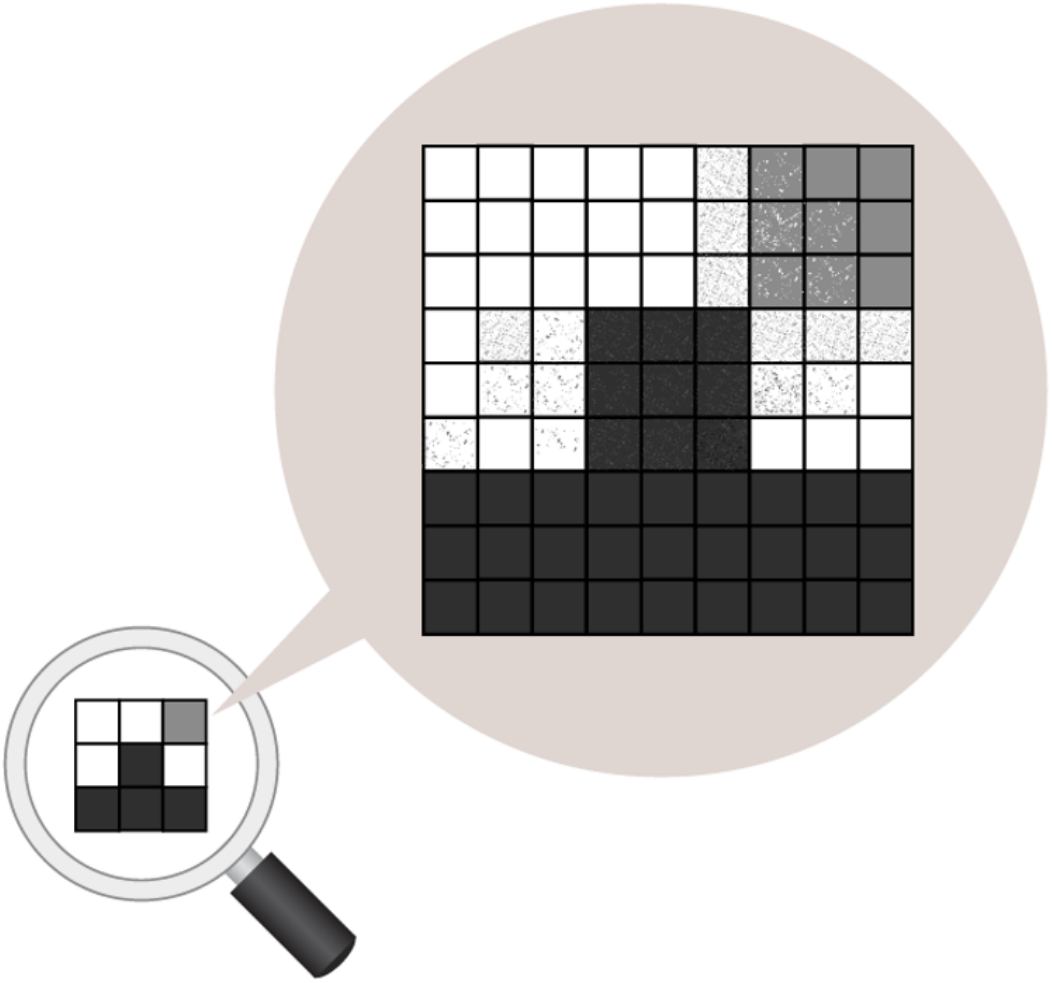


Cai as a Cartesian treats its sensory experiences as though they exist in a separate world.

Cartesian Cai recognizes that its two universes, its sensory experiences and hypothesized environment, can interact. But it thinks they can only do so via a narrow

range of stable pathways. No actual agent's mind-matter connections can be that simple and uniform.

If Cai were a robot in a world resembling its model, it would *itself* be a complex pattern of tiles. To form accurate predictions, it would need to have self-models and bridge hypotheses that were more sophisticated than any I've considered so far. Humans are the same way: No bridge hypothesis explaining the physical conditions for subjective experience will ever fit on a T-shirt.



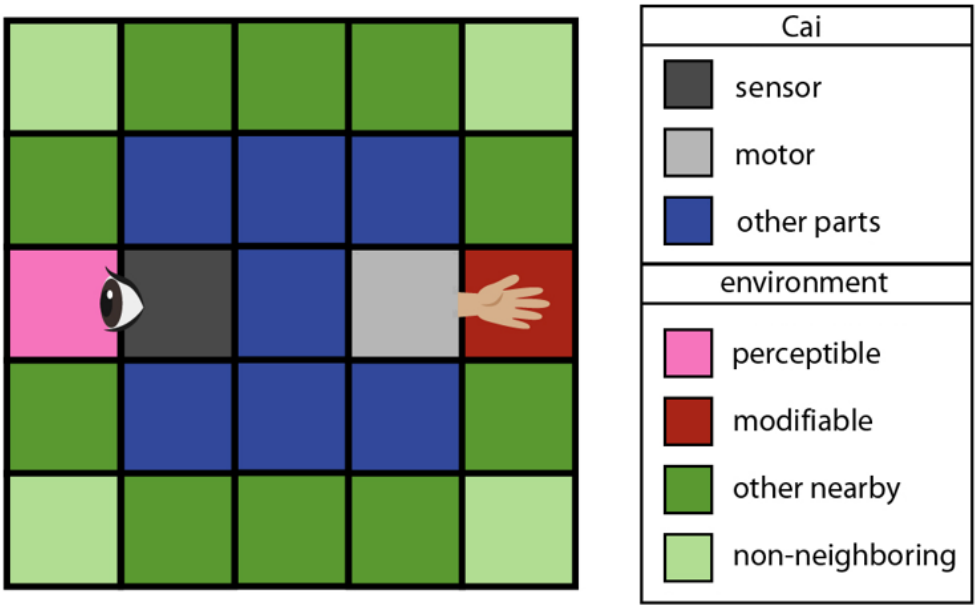
Cai's world divided up into a 9x9 grid. Cai is the central 3x3 grid. Barely visible: Complex computations like Cai's reasoning are possible in this world because they're implemented by even finer tile patterns at smaller scales.

Changing Cai's tiles' states — from black to white, for example — could have a large impact on its computations, analogous to changing a human brain from solid to gaseous. But if an agent's hypotheses are all shaped like the cybernetic agent model, 'my input/output algorithm is replaced by a dust cloud' won't be in the hypothesis space.

If you programmed something to thinks like Cartesian Cai, it might decide that its sequence of visual experiences will persist even if the tiles forming its brain completely change state. It wouldn't be able to entertain thoughts like 'if Cai performs self-modification #381, Cai will experience its environment as smells rather than colors' or 'if Cai falls into a volcano, Cai gets destroyed'. No pattern of perceived colors is identical to a perceived smell, or to the absence of perception.

To form naturalistic self-models and world-models, Cai needs hypotheses that look less like conversations between independent programs, and more like worlds in which it is a fairly ordinary subprocess, governed by the same general patterns. It needs to form and privilege physical hypotheses under which it has parts, as well as bridge hypotheses under which those parts correspond in plausible ways to its high-level computational states.

Cai wouldn't need a *complete* self-model in order to recognize general facts about its subsystems. Suppose, for instance, that Cai has just one sensor, on its left side, and a motor on its right side. Cai might recognize that the motor and sensor regions of its body correspond to its introspectible decisions and perceptions, respectively.

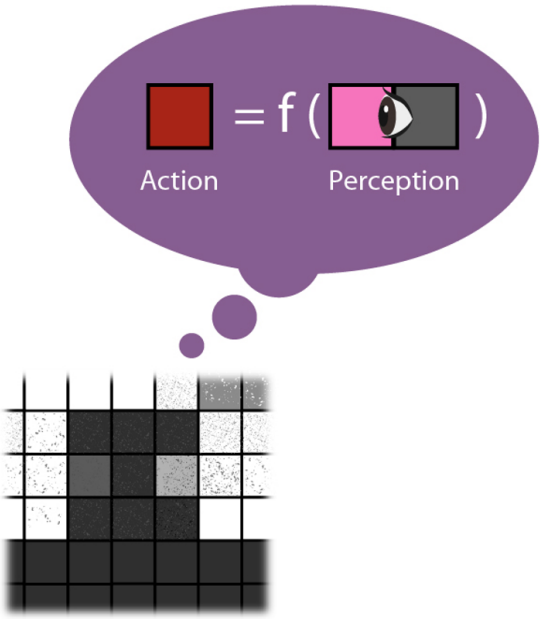


A naturalized agent can recognize that it has physical parts with varying functions. Cai's top and bottom lack sensors and motors altogether, making it clearer that Cai's environment can impact Cai by entirely non-sensory means.

We care about Cai's models because we want to use Cai to modify its environment. For example, we may want Cai to convert as much of its environment as possible into grey tiles. Our interest is then in the algorithm that reliably outputs maximally greyifying actions when handed perceptual data.

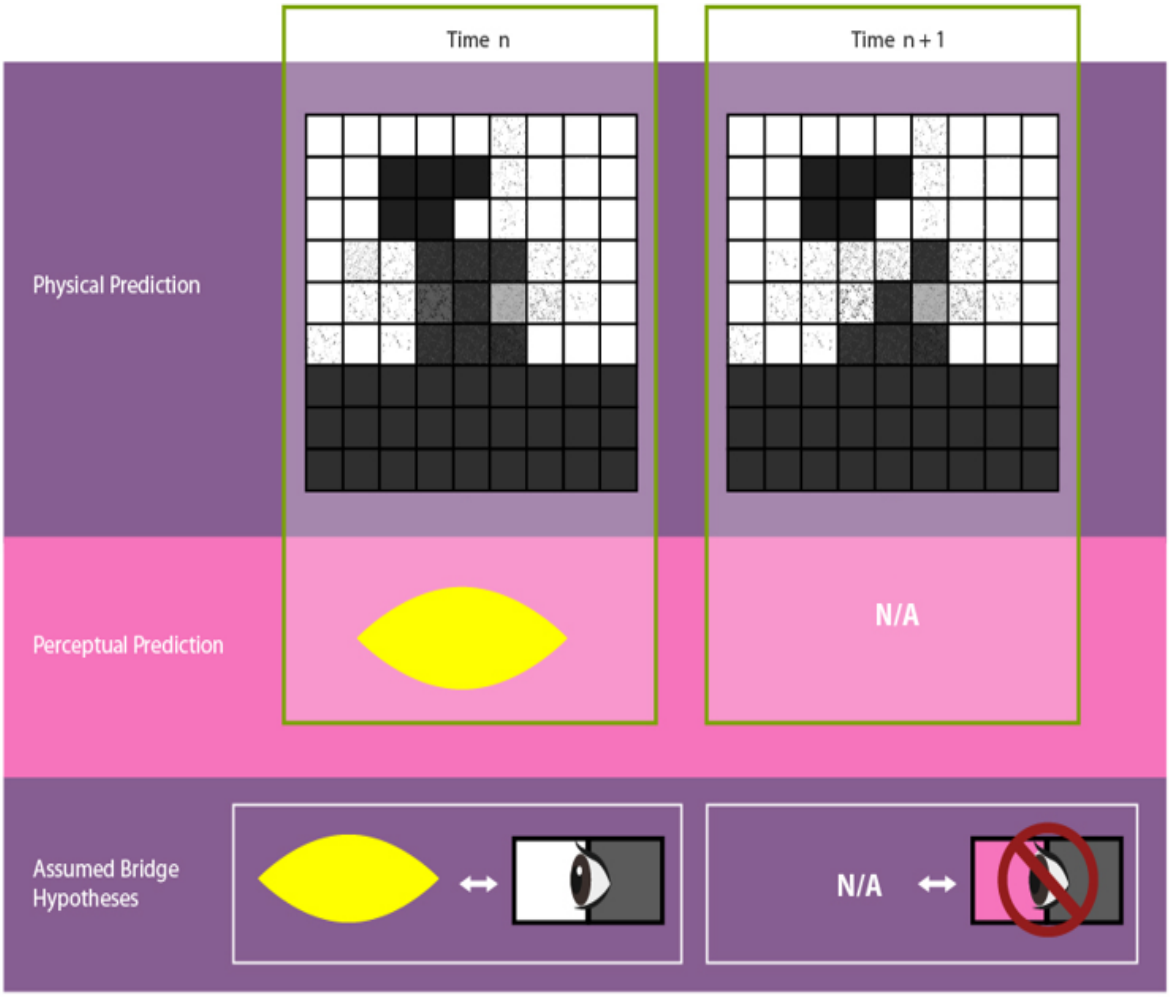
If Cai is able to form sophisticated self-models, then Cai can recognize that it's a grey tile maximizer. Since it wants there to be more grey tiles, it also wants to make sure that it continues to exist, provided it believes that it's better than chance at pursuing its goals.

More specifically, Naturalized Cai can recognize that its actions are some black-box function of its perceptual computations. Since it has a bridge hypothesis linking its perceptions to its middle-left tile, it will then reason that it should *preserve* its sensory hardware. Cai's self-model tells it that if its sensor fails, then its actions will be based on beliefs that are much less correlated with the environment. And its self-model tells it that if its actions are poorly calibrated, then there will be fewer grey tiles in the universe. Which is bad.



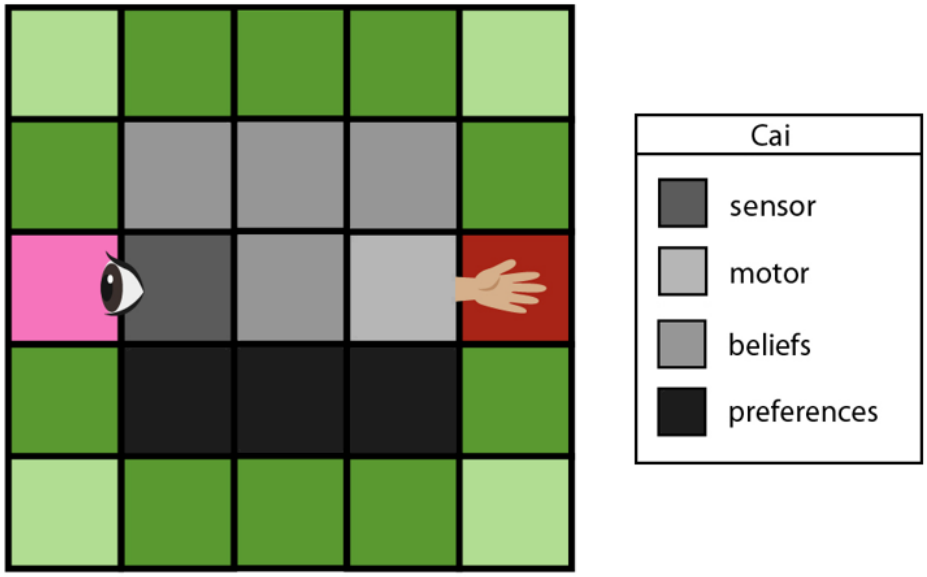
A naturalistic version of Cai can reason intelligently from the knowledge that its actions (motor output) depend on a specific part of its body that's responsible for perception (environmental input).

A physical Cai might need to foresee scenarios like 'an anvil crashes into my head and destroys me', and assign probability mass to them. Bridge hypotheses expressive enough to consider that possibility would not just relate experiences to environmental or hardware states; they would also recognize that the agent's experiences can be absent altogether.



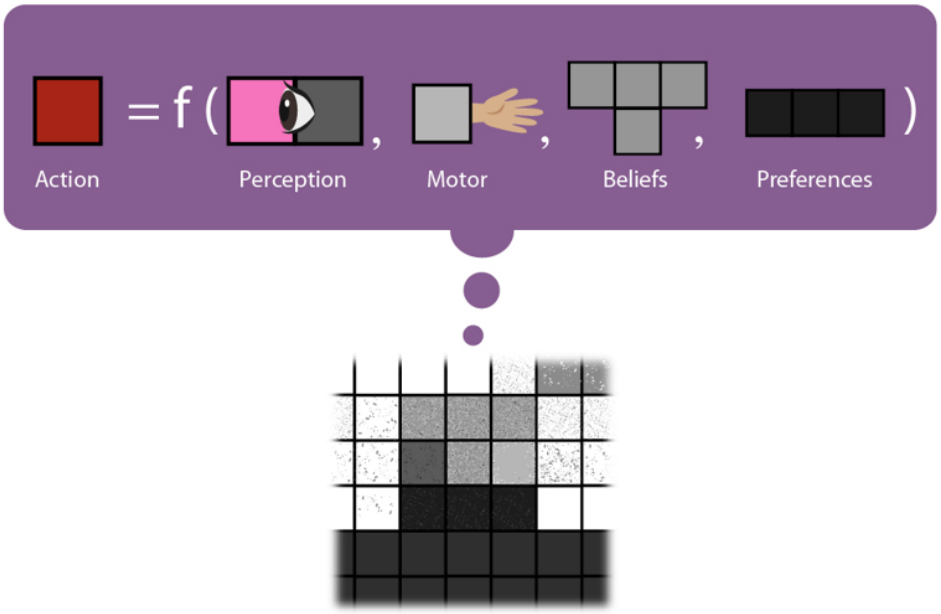
An anvil can destroy Cai's perceptual hardware by crashing into it. A Cartesian might not worry about this eventuality, expecting its experience to persist after its body is smashed. But a naturalized reasoner will form hypotheses like the above, on which its sequence of color experiences suddenly terminates when its sensors are destroyed.

This point generalizes to other ways Cai might self-modify, and to other things Cai might alter about itself. For example, Cai might learn that other portions of its brain correspond to its hypotheses and desires.



Another very simple model of how different physical structures are associated with different computational patterns.

This allows Cai to recognize that its goals depend on the proper functioning of *many* of its hardware components. If Cai believes that its actions depend on its brain's goal unit's working a specific way, then it will avoid taking pills that foreseeably change its goal unit. If Cai's causal model tells it that agents like it stop exhibiting future-steering behaviors when they self-modify to have mad priors, then it won't self-modify to acquire mad priors. And so on.



If Cai's motor fails, its effect on the world can change as a result. The same is true if its hardware is modified in ways that change its thoughts, or its preferences (i.e., the thing linking its conclusions to its motor).

Once Cai recognizes that its brain needs to work in a very specific way for its goals to be achieved, its preferences can take its physical state into account in sensible ways, without our needing to hand-code Cai at the outset to have the right beliefs or preferences over every individual thing that could change in its brain.

Just the opposite is true for Cartesians. Since they can't form hypotheses like 'my tape heads will stop computing digits if I disassemble them', they can only intelligently navigate such risks if they've been hand-coded in advance to avoid perceptual experiences the programmer thought would correlate with such dangers.

In other words, even though all of this is still highly informal, there's already some cause to think that a reasoning pattern like Naturalized Cai can generalize in ways that Cartesians can't. The programmers don't need to know *everything* about Cai's physical state, or anticipate *everything* about what future changes Cai might undergo, if Cai's epistemology allows it to easily form accurate reductive beliefs and behave accordingly. An agent like this might be adaptive and self-correcting in very novel circumstances, leaving more wiggle room for programmers to make human mistakes.

Bridging maps of worlds and maps of minds

Solomonoff-style dualists have alien blind spots that lead them to neglect the possibility that some hardware state is equivalent to some introspected computation '000110'. TALE-SPIN-like AIs, on the other hand, have blind spots that lead to mistakes like trying to figure out the angular momentum of '000110'.

A naturalized agent doesn't try to do away with the data/hypothesis type distinction and acquire a typology as simple as TALE-SPIN's. Rather, it tries to tightly interconnect its types using bridges. Naturalizing induction is about combining the dualist's useful map/territory distinction with a more sophisticated metaphysical monism than TALE-SPIN exhibits, resulting in a *reductive* monist AI.[7]

Alice's simple fixed bridge axiom, {environmental output 0 ↔ perceptual input 0, environmental output 1 ↔ perceptual input 1}, is inadequate for physically embodied agents. And the problem isn't *just* that Alice lacks other bridge rules and can't weigh evidence for or against each one. Bridge hypotheses are a step in the right direction, but they need to be diverse enough to express a variety of correlations between the agent's sensory experiences and the physical world, and they need a sensible prior. An agent that only considers bridge hypotheses compatible with the cybernetic agent model will falter whenever it and the environment interact in ways that look nothing like exchanging sensory bits.

With the help of an inductive algorithm that uses bridge hypotheses to relate sensory data to a continuous physical universe, we can avoid making our AIs Cartesians. This will make their epistemologies much more secure. It will also make it possible for them to want things to be true about the physical universe, not just about the particular sensory experiences they encounter. Actually writing a program that does all this is an OPFAI. Even formalizing how bridge hypotheses ought to work in principle is an OPFAI.

In my next post, I'll move away from toy models and discuss AIXI, Hutter's optimality definition for cybernetic agents. In asking whether the *best* Cartesian can overcome the difficulties I've described, we'll get a clearer sense of why Solomonoff inductors aren't reflective and reductive enough to predict drastic changes to their sense-input-to-motor-output relation — and why they *can't be* that reflective and reductive — and why this matters.

Notes

[1] Meehan (1977). Colin Allen first introduced me to this story. Dennett discusses it as well. ↩

[2] E.g., Durand, Muchnik, Ushakov & Vereshchagin (2004), Epstein & Betke (2011), Legg & Veness (2013), Solomonoff (2011). Hutter (2005) uses the term "cybernetic agent model" to emphasize the parallelism between his Turing machine circuit and control theory's cybernetic systems. ↩ ↩

[3] One simple representation would be: Program Alice to write to her work tape, on round one, 0010 (standing for 'if I output 0, Everett outputs 0; if I output 1, Everett outputs 0'). Ditto for the other three hypotheses, 0111, 0011, and 0110. Then write the hypothesis' probability in binary (initially 25%, represented '11001') to the right of each, and program Alice to edit this number as she receives new evidence. Since the first and third digit stay the same, we can simplify the hypotheses' encoding to 00, 11, 01, 10. Indeed, if the hypotheses remain the same over time there's no reason to visibly distinguish them in the work tape at all, when we can instead just program Alice to use the left-to-right ordering of the four probabilities to distinguish the hypotheses. ↩

[4] To the extent our universe *perfectly* resembles any mathematical structure, it's much more likely to do so at the level of gluons and mesons than at the level of medium-sized dry goods. The resemblance of apples to natural numbers is much more approximate. Two apples and three apples generally make five apples, but when you start cutting up or pulverizing or genetically altering apples, you may find that other mathematical models do a superior job of predicting the apples' behavior. It seems likely that the only perfectly general and faithful mathematical representation of apples will be some drastically large and unwieldy physics equation.

Ditto for machines. It's sometimes possible to build a physical machine that closely mimics a given Turing machine — but only 'closely', as Turing machines have unboundedly large tapes. And although any halting Turing machine can in principle be simulated with a bounded tape (Cockshott & Michaelson (2007)), nearly all Turing machine programs are too large to even be approximated by any physical process.

All physical machines structurally resemble Turing machines in ways that allow us to draw productive inferences from the one group to the other. See Piccinini's (2011) discussion of the physical Church-Turing thesis. But, for all that, the concrete machine and the abstract one remain distinct. ↩

[5] Descartes (1641): "[A]lthough I certainly do possess a body with which I am very closely conjoined; nevertheless, because, on the one hand, I have a clear and distinct idea of myself, in as far as I am only a thinking and unextended thing, and as, on the

other hand, I possess a distinct idea of body, in as far as it is only an extended and unthinking thing, it is certain that I (that is, my mind, by which I am what I am) am entirely and truly distinct from my body, and may exist without it."

From this it's clear that Descartes also believed that the mind can exist without the body. This interestingly parallels the anvil problem, which I'll discuss more in my next post. However, I don't build immortality into my definition of 'Cartesianism'. Not all agents that act as though there is a Cartesian barrier between their thoughts and the world think that their experiences are future-eternal. I'm taking care not to conflate Cartesianism with the anvil problem because the formalism I'll discuss next time, AIXI, does face both of them. Though the problems are logically distinct, it's true that a naturalized reasoning method would be much less likely to face the anvil problem. ↩

[6] This isn't to say that a Solomonoff inductor would need to be conscious in anything like the way humans are conscious. It can be fruitful to point to similarities between the reasoning patterns of humans and unconscious processes. Indeed, this already happens when we speak of unconscious mental processes within humans.

Parting ways with Descartes (cf. Kirk (2012)), many present-day dualists would in fact go even further than reductionists in allowing for structural similarities between conscious and unconscious processes, treating all cognitive or functional mental states as (in theory) realizable without consciousness. E.g., Chalmers (1996): "Although consciousness is a feature of the world that we would not predict from the physical facts, the things we say about consciousness are a garden-variety cognitive phenomenon. Somebody who knew enough about cognitive structure would immediately be able to predict the likelihood of utterances such as 'I feel conscious, in a way that no physical object could be,' or even Descartes's 'Cogito ergo sum.' In principle, some reductive explanation in terms of internal processes should render claims about consciousness no more deeply surprising than any other aspect of behavior." ↩

[7] And since we happen to live in a world made of physics, the kind of monist we want in practice is a reductive *physicalist* AI. We want a 'physicalist' as opposed to a reductive monist that thinks everything is made of monads, or abstract objects, or morality fluid, or what-have-you. ↩

References

• Chalmers (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

• Cockshott & Michaelson (2007). Are there new models of computation? Reply to Wegner and Eberbach. *The Computer Journal, 50*: 232-247.

• Descartes (1641). *Meditations on first philosophy, in which the existence of God and the immortality of the soul are demonstrated*.

• Durand, Muchnik, Ushakov & Vereshchagin (2004). Ecological Turing machines. *Lecture Notes in Computer Science, 3142*: 457-468.

• Epstein & Betke (2011). An information-theoretic representation of agent dynamics as set intersections. *Lecture Notes in Computer Science, 6830*: 72-81.

• Hutter (2005). *Universal Artificial Intelligence: Sequence Decisions Based on Algorithmic Probability*. Springer.

• Kirk (2012). Zombies. In Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

• Legg & Veness (2013). An approximation of the Universal Intelligence Measure. *Lecture Notes in Computer Science, 7070*: 236-249.

• Meehan (1977). TALE-SPIN, an interactive program that writes stories. *Proceedings of the 5th International Joint Conference on Artificial Intelligence*: 91-98.

• Piccinini (2011). The physical Church-Turing thesis: Modest or bold? *British Journal for the Philosophy of Science, 62*: 733-769.

• Russell & Norvig (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

• Solomonoff (2011). Algorithmic probability — its discovery — its properties and application to Strong AI. In Zenil (ed.), *Randomness Through Computation: Some Answers, More Questions* (pp. 149-157).

Can We Do Without Bridge Hypotheses?

Followup to: Building Phenomenological Bridges, Reductionism

Bridge hypotheses are extremely awkward. It's risky to draw permanent artificial lines between categories of hypothesis ('physical' vs. 'bridge'). We might not give the right complexity penalties to one kind of hypothesis relative to the other. Or we might implement a sensible framework for bridge hypotheses in one kind of brain that fails to predict the radically new phenomenology that results from expanding one's visual cortex onto new hardware.

We'd have to hope that it makes sense to talk about 'correct' bridging rules (correctly relating a hypothesis about external stimuli or about transistors composing yourself, to which settings are in fact the ones you call 'green'), even though they're quite different from ordinary physical descriptions of the world. And, since fully general and error-free knowledge of the phenomenologies of possible agents will probably not be available to a seed AGI or to its programmers, we'd have to hope that it's possible to build a self-modifying inductor robust enough that mistaken bridge predictions would just result in a quick Bayesian update towards better ideas. It's definitely a dangling thread.

Why, then, can't we do without them? Maybe they're a handy heuristic for agents with incomplete knowledge — but can they truly *never* be eliminated?

The notion of an irreducible divide between an AI's subjective sensations and its models of the objective world may sound suspiciously dualistic. If we live in a purely physical world, then why shouldn't a purely physical agent, once it’s come to a complete understanding of itself and the world, be able to dispense with explicit bridges? These are, after all, the agent's *beliefs* that we're talking about. In the limit, intuitively, accurate beliefs should just look like the world. So shouldn't the agent's phenomenological self-models eventually end up collapsing into its physical world-models — dispensing with a metaphysically basic self/world distinction?[1]

Yes and no. When humans first began hypothesizing about the relationship between mind and matter, the former domain did not appear to be reducible to the latter. A number of philosophers concluded from this that there was a deep metaphysical divide between the two. But as the sciences of mind began to erode that belief in mind-matter dualism, they didn't eliminate the conceptual, linguistic, or intuitive distinctness of our mental and physical models. It may well be that we'll never abandon an intentional stance toward many phenomena, even once we've fully reduced them to their physical, biological, or computational underpinnings. Models of different levels can remain useful even once we've recognized that they co-refer.

In the case of an artificial scientist, beliefs in a fundamental sensation-v.-world dichotomy may dissolve even if the agent retains a useful conceptual distinction between its perceptual stream and the rest of the world. A lawful, unified physics need not be best modeled by agents with only a single world-modeling subprocess. 'There

is one universe' doesn't imply 'one eye is optimal for viewing the universe'; 'there is one Earth' doesn't imply 'one leg is optimal for walking it'. The cases seem different chiefly because the leg/ground distinction is easier for humans to keep straight than the map/territory distinction.

Empirical reasoning requires a representational process that produces updates, and another representational process that gets updated. Eliminate the latter, and gone is the AI's memory and expectation. (Imagine Cai experiencing its sequence of colors forever without considering any states of affairs they predict.) Eliminate the former, and the AGI has nothing but its frozen memories. (Imagine Cai without any sensory input, just a floating array of static world-models.) Keep both and eliminate bridging, and Cai painstakingly collects its visual data only to throw it all away; it has beliefs, but it never updates them.

Can we replace perceptions and expectations with a single kind-of-perceptiony kind-of-expectationish epistemic process, in a way that obviates any need for bridge hypotheses?

Maybe, but I don't know what that would look like. An agent's perceptions and its hypotheses are of different types, just by virtue of having distinct functions; and its meta-representations must portray them as such, lest its metacognitive reasoning fall into systemic error. Striving mightily to conflate the two may not make any more sense than striving to get an agent to smell colors or taste sounds.[2]

The only candidate I know of for a framework that may sidestep this distinction without thereby catching fire is Updateless Decision Theory, which was brought up by Jim Babcock, Vladimir Slepnev, and Wei Dei. UDT eliminates the need for bridge hypotheses in a particularly bold way, by doing away with updatable hypotheses altogether.

I don't understand UDT well enough to say how it bears on the problem of naturalizing induction, but I may return to this point when I have a better grasp on it. If UDT turns out to solve or dissolve the problem, it will be especially useful to have on hand a particular reductionism-related problem that afflicts other kinds of agents and is solved by UDT. This will be valuable even if UDT has other features that are undesirable enough to force us to come up with alternative solutions to naturalized induction.

For now, I'll just make a general point: It's usually good policy for an AGI to think like reality; but if an introspectible distinction between updatable information and update-causing information *is* useful for real-world inductors, then we shouldn't strip all traces of it from artificial reasoners, for much the same reason we shouldn't reduce our sensory apparatuses to a single modality in an attempt to ape the unity of our world's dynamics. Reductionism restricts what we can rationally believe about the territory, but it doesn't restrict the idiom of our maps.

---

[1] This is close to the worry Alex Flint raised, though our main concern is with the agent's ability to reduce its own mental types, since this is a less avoidable problem than a third party trying to do the same.

[2] The analogy to sensory modality is especially apt given that phenomenological bridge hypotheses can link sensory channels instead of linking a sensory channel to a hypothesized physical state. For instance, '*I see yellow whenever I taste isoamyl acetate*' can function as a bridge between sensations an agent types as 'vision' and sensations an agent types as 'taste'.

Solomonoff Cartesianism

Followup to: Bridge Collapse; An Intuitive Explanation of Solomonoff Induction; Reductionism

Summary: If you want to predict arbitrary computable patterns of data, Solomonoff induction is the optimal way to go about it — provided that you're an eternal transcendent hypercomputer. A real-world AGI, however, won't be immortal and unchanging. It will need to form hypotheses about its own physical state, including predictions about possible upgrades or damage to its hardware; and it will need bridge hypotheses linking its hardware states to its software states. As such, the project of building an AGI demands that we come up with a new formalism for constructing (and allocating prior probabilities to) hypotheses. It will not involve just building increasingly good computable approximations of AIXI.

Solomonoff induction has been cited repeatedly as the theoretical gold standard for predicting computable sequences of observations.[1] As Hutter, Legg, and Vitanyi (2007) put it:

> Solomonoff's inductive inference system will learn to correctly predict any computable sequence with only the absolute minimum amount of data. It would thus, in some sense, be the perfect universal prediction algorithm, if only it were computable.

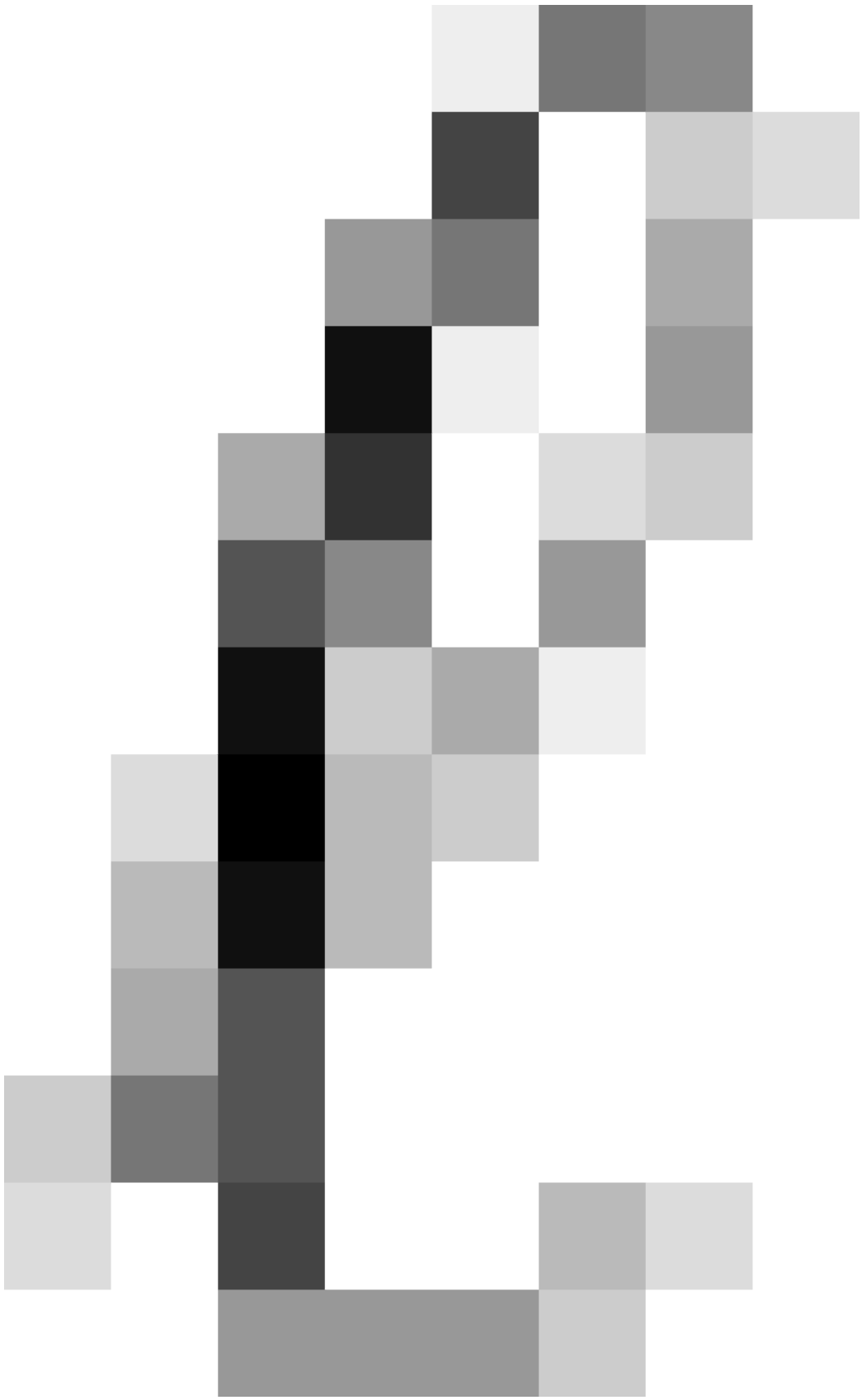
Perhaps you've been handed the beginning of a sequence like 1, 2, 4, 8… and you want to predict what the next number will be. Perhaps you've paused a movie, and are trying to guess what the next frame will look like. Or perhaps you've read the first half of an article on the Algerian Civil War, and you want to know how likely it is that the second half describes a decrease in GDP. Since all of the information in these scenarios can be represented as patterns of numbers, they can all be treated as rule-governed sequences like the 1, 2, 4, 8… case. Complicated sequences, but sequences all the same.

It's been argued that in all of these cases, one unique idealization predicts what comes next better than any computable method: Solomonoff induction. No matter how limited your knowledge is, or how wide the space of computable rules that could be responsible for your observations, the ideal answer is always the same: Solomonoff induction.

Solomonoff induction has only a few components. It has one free parameter, a choice of universal Turing machine. Once we specify a Turing machine, that gives us a fixed encoding for the set of all possible programs that print a sequence of 0s and 1s. Since every program has a specification, we call the number of bits in the program's specification its "complexity"; the shorter the program's code, the simpler we say it is.

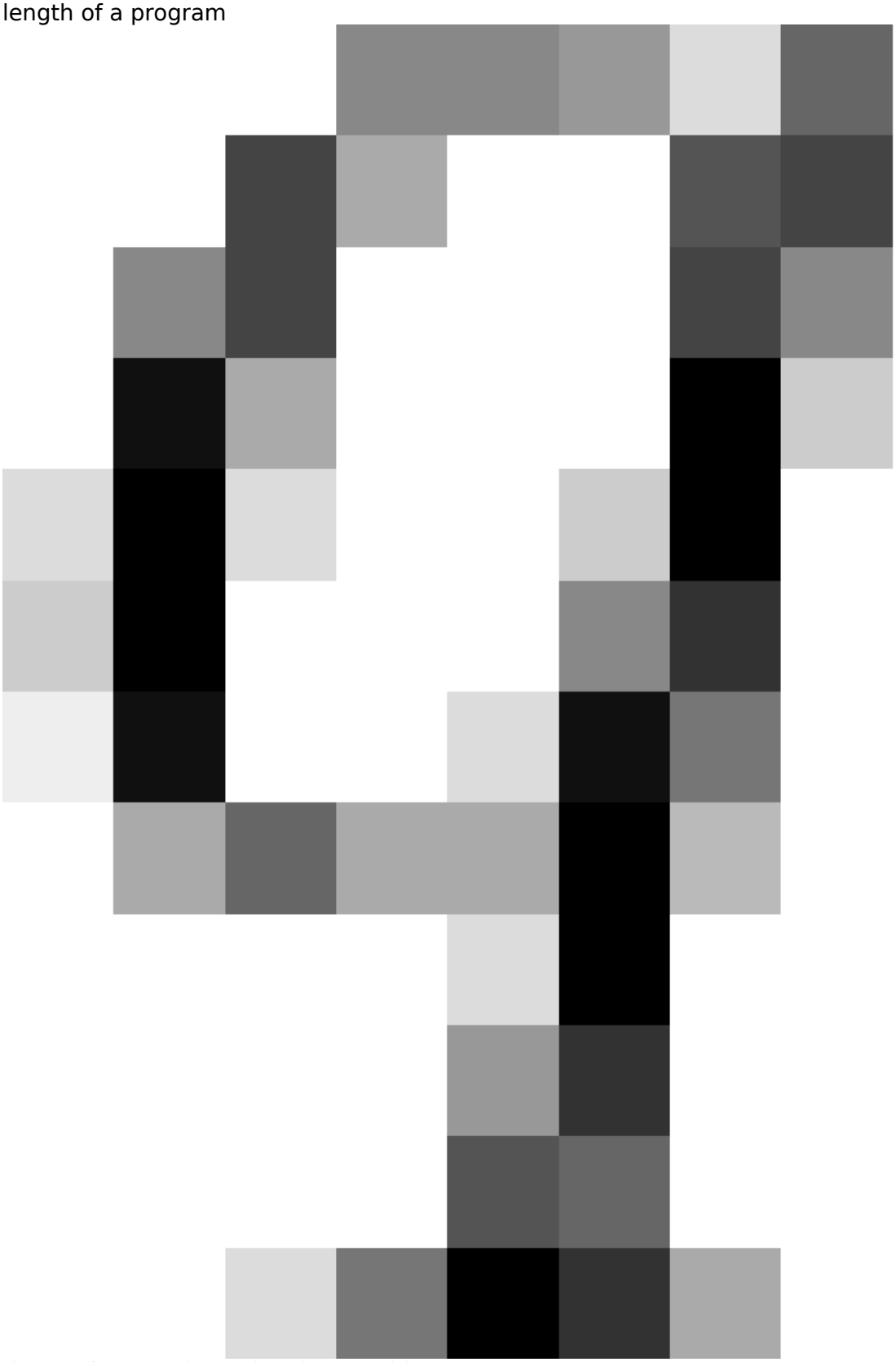
Solomonoff induction takes this infinitely large bundle of programs and assigns each one a prior probability proportional to its simplicity. Every time the program requires one more bit, its prior probability goes down by a factor of 2, since there are then twice as many possible computer programs that complicated. This ensures the sum over all programs' prior probabilities equals 1, even though the number of programs is infinite.[2]

The imaginary inductor is then fed a sequence of 0s and 1s, and with each new bit it updates using Bayes' rule to promote programs whose outputs match the observed sequence. So, where

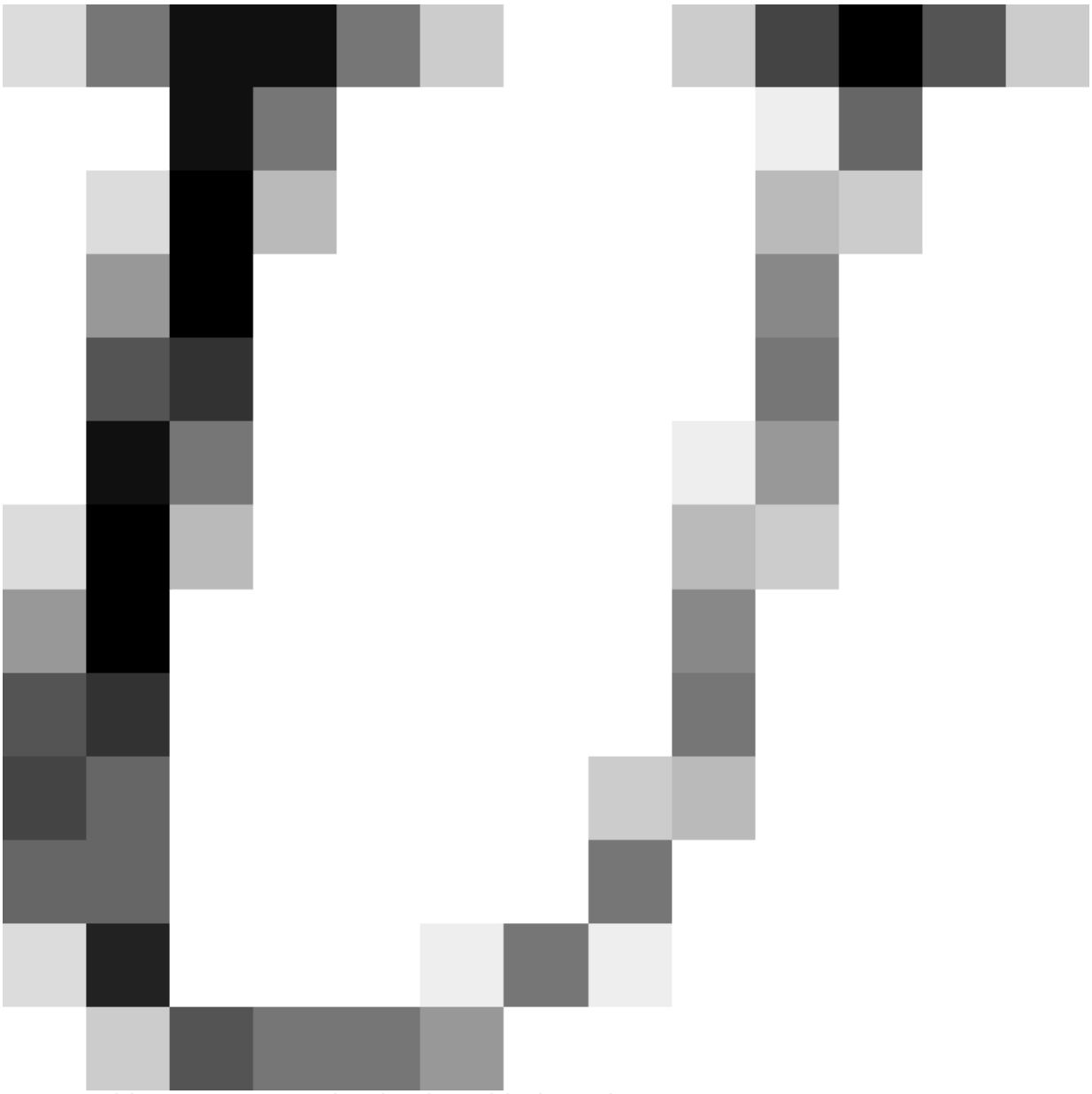


is the

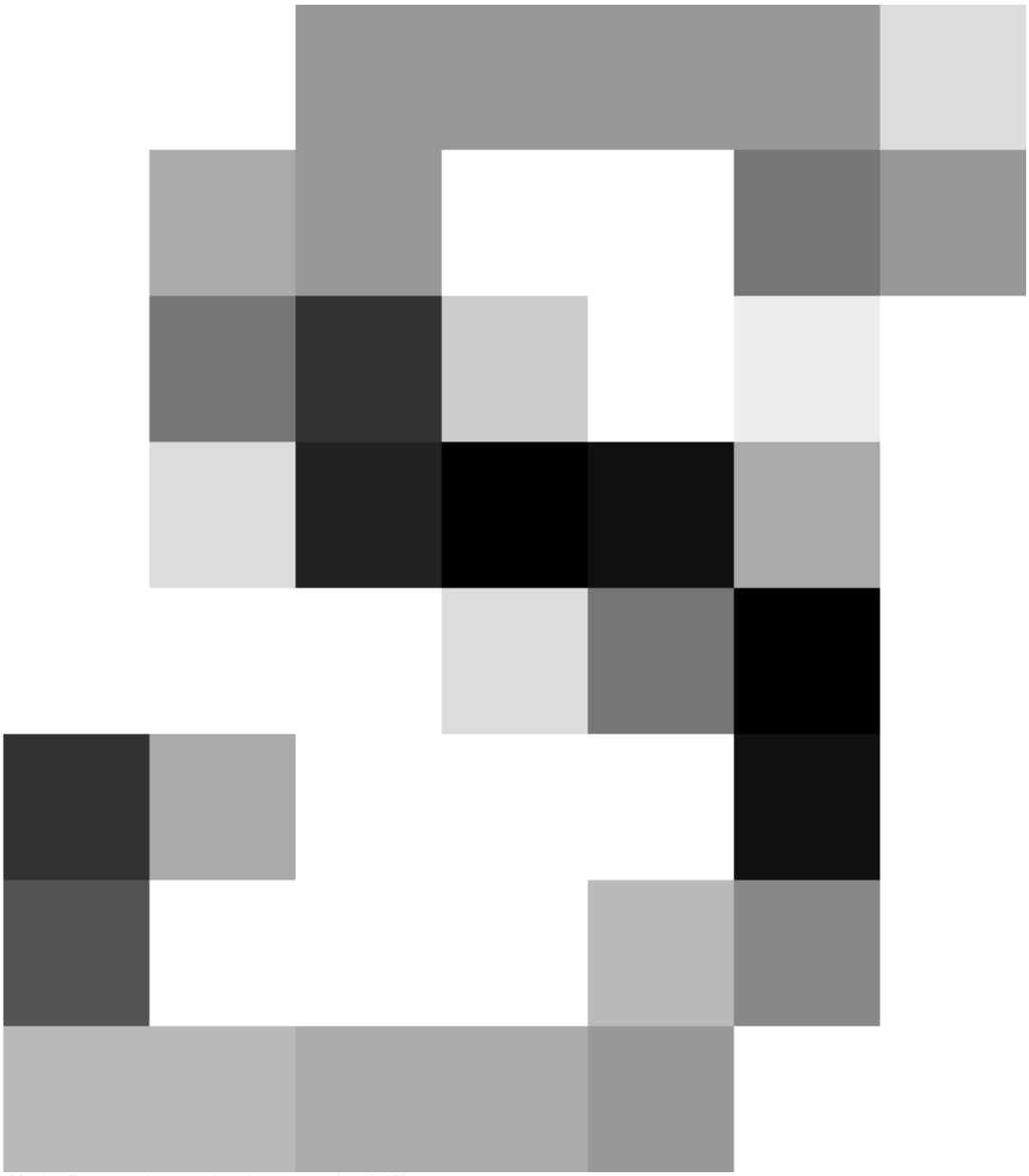
length of a program



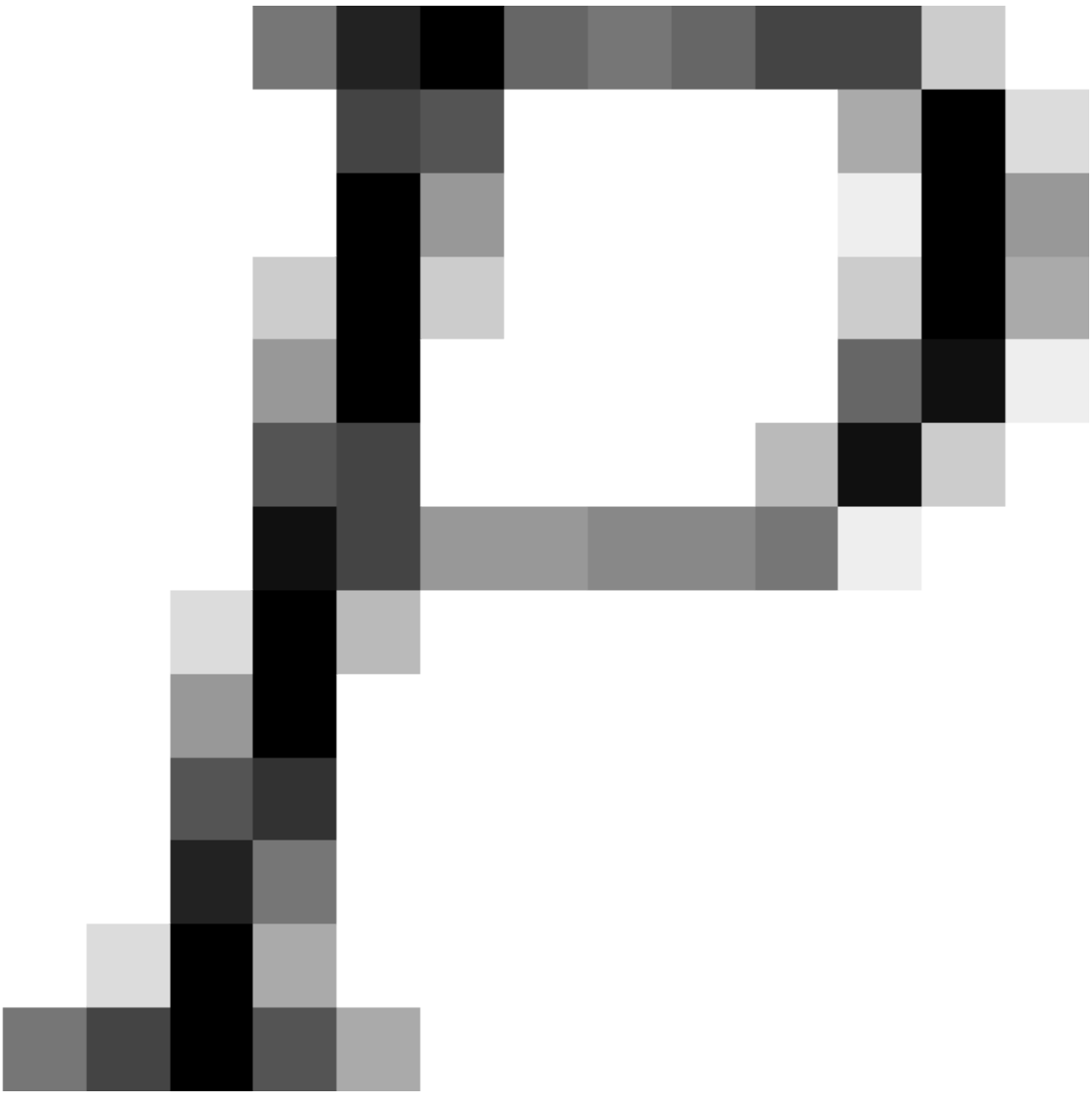
that makes a universal Turing machine



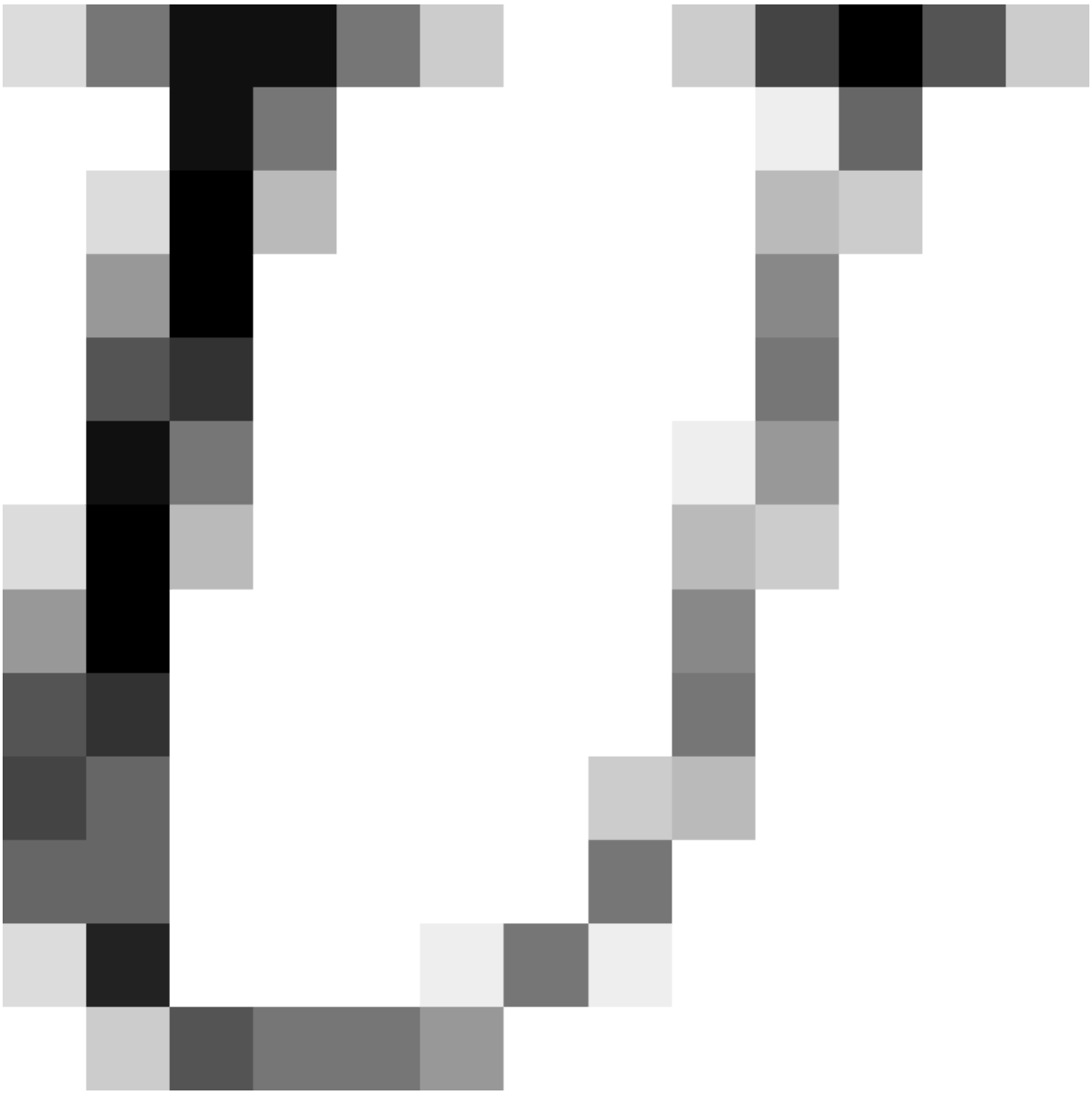
output a binary sequence that begins with the string



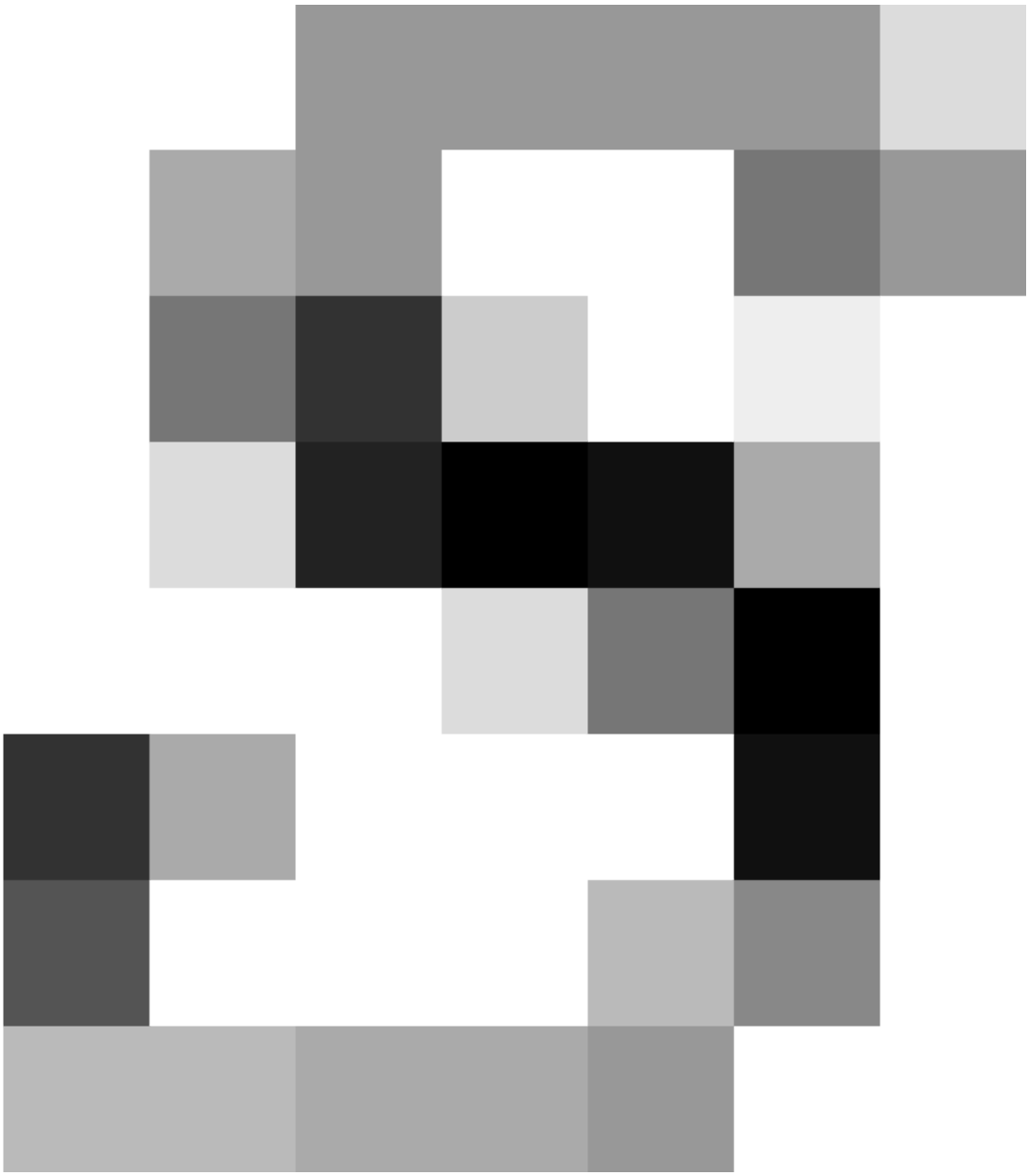
, Solomonoff defines the relative probability



that



outputs



:

$$P^U(s) \quad := \quad \sum_{U(q)=s\ldots} 2^{-\ell(q)}$$

Solomonoff induction isn't computable, but it's been singled out as the unbeatable formal predictor of computable sequences.[3] All computable rules for generating sequences are somewhere within Solomonoff's gargantuan bundle of programs. This includes all rules that a human brain could use. If the rule that best matches the observations is 1000 bits large, it will take at most 1000 bits of evidence — 1000 bits worth of predictions made better than any other rule — for that rule to be promoted to the top of consideration. Solomonoff's claim to being an optimal ideal rule rests on the fact that it never does worse than any computable rule (including you!) by more than a fixed amount.[4]

Who cares, if we can't build the thing?

Encouraged by Solomonoff inductors' optimality properties, some have suggested that building a working AGI calls for little more than finding out which computable algorithm comes as close to Solomonoff induction as possible given resource constraints, and supplying an adequate learning environment and decision criterion.[5]

Eliezer Yudkowsky thinks that these attempts to approximate Solomonoff are a dead end. Much of the difficulty of intelligence rests on computing things cheaply, and Yudkowsky doesn't think that the kind of search these algorithms are doing will zero in on cheap ways to reason. There are practical lessons to be learned from Solomonoff induction, but the particular kind of optimality Solomonoff induction exhibits depends in important ways on its computational unfeasibility,[4] which makes it unlikely that Solomonoff imitators will ever be efficient reasoners.

Why, then, should Solomonoff induction interest us? If we can't execute it, and we can't design useful AGIs by *directly* emulating it, then what's it good for?

My answer is that if Solomonoff induction *would* deliver flawless answers, could we but run it, then it has a claim to being an ideal mirror to which we can hold up instances of human and artificial inductive reasoning.

In From Philosophy to Math to Engineering, Muehlhauser talks about how ideas often progress from productive but informal ruminations ('philosophy'), to rigorously specified idealizations ('mathematics'), to functioning technologies ('engineering').

Solomonoff inductors fall into the second category, 'mathematics': We could never build them, but thinking in terms of them can give us useful insights and point us in the right direction. For example, Solomonoff's ideal can remind us that privileging simple hypotheses isn't just a vague human fancy or quirk; it has formalizations with situation-invariant advantages we can state with complete precision. It matters that we can pinpoint the sense in which a lengthy physical or meteorological account of lightning is simpler than 'Thor did it', and it matters that we can cite reasons for giving more credence to hypotheses when they have that kind of simplicity.[6]

Bayesian updating is usually computationally intractable, but as an ideal it gives us a simple, unified explanation for a wealth of observed epistemic practices: They share structure in common with a perfect Bayesian process. Similarly, optimality proofs for Solomonoff's prior can yield explanations for why various real-world processes that privilege different notions of simplicity succeed or fail.

Though Solomonoff induction is uncomputable, if it is truly the *optimal* reasoning method, then we have found at least one clear ideal we can use to compare the merits

of real-world algorithms for automating scientific reasoning.[1] But that 'if' is crucial. I haven't yet spoken to the question of whether Solomonoff induction *is* a good background epistemology, analogous to Bayesianism.

Where Solomonoff induction goes wrong

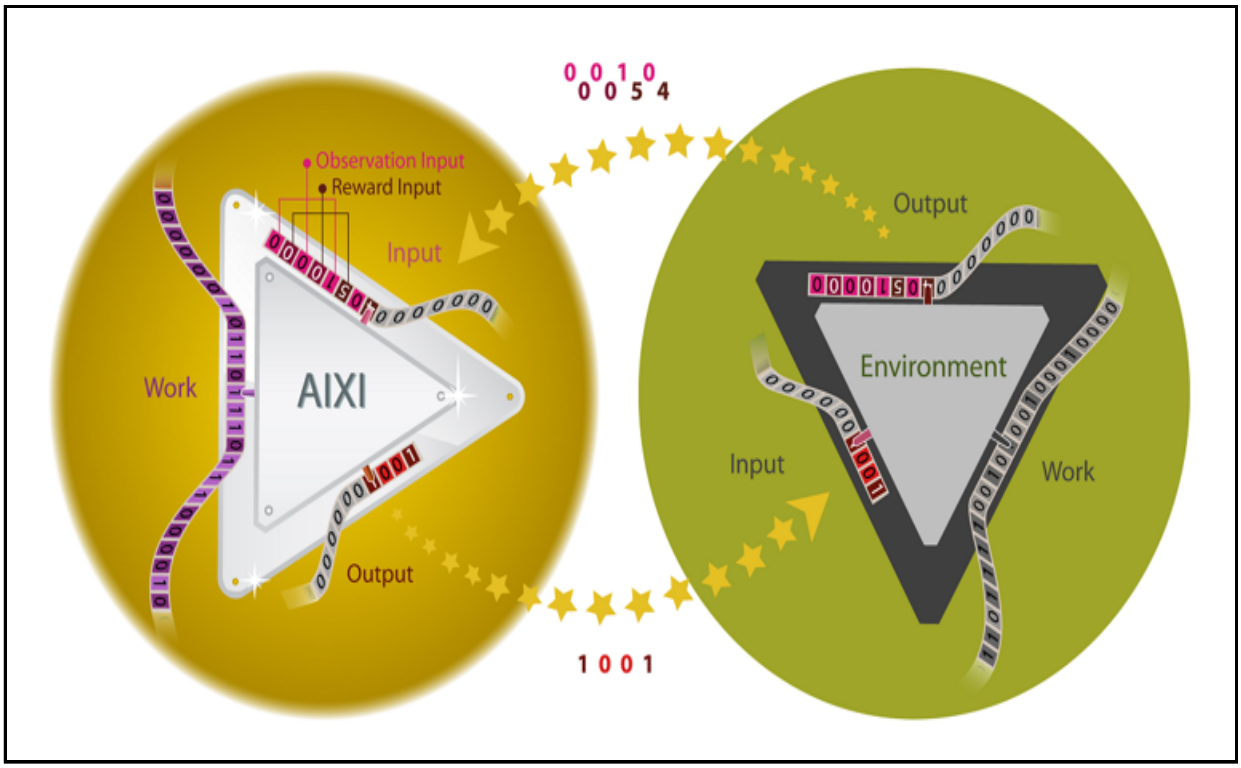
My claim will be that, computational difficulties aside, Solomonoff induction is not an adequate mathematical definition of ideal inductive reasoning.

What follows will be a first-pass problem statement, giving background on why naturalizing induction may require us to construct an entirely new, non-Solomonoff-based paradigm for intelligence. This is preliminary; formalizations of the problem will need to wait until a second pass, and we don't have a fleshed-out solution to offer, though we can gesture toward some possible angles of attack. But I can begin to illustrate here why Solomonoff inductors have serious limitations that can't be chalked up to their uncomputability.

In Bridge Collapse, I defined **Cartesianism** as the belief that one's internal computations cannot be located in the world. For a Cartesian, sensory experiences are fundamentally different in type from the atoms of the physical world. The two can causally interact, but we can never completely reduce the former to the latter.

Solomonoff inductors differ greatly from human reasoners, yet they are recognizably *Cartesian*. Broadly dualistic patterns of reasoning crop up in some decidedly inhuman algorithms. (Admittedly, algorithms invented by humans.)

This core limitation of Solomonoff induction can be seen most clearly when it results in an AI that not only *thinks* in bizarre ways, but also acts accordingly. I'll focus on **AIXI**, Marcus Hutter's hypothetical design for a Solomonoff inductor hooked up to an expected reward signal maximizer.

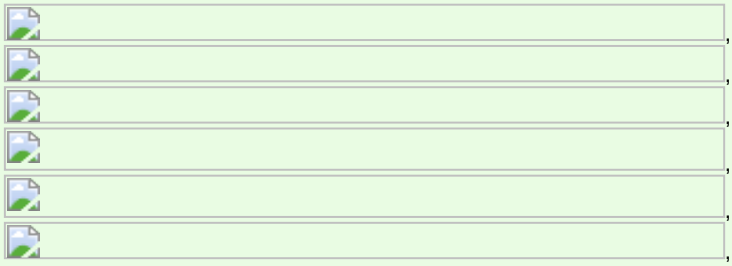


Hutter's [cybernetic agent model] of AIXI. AIXI outputs whichever actions it expects to cause an environmental Turing machine to output rewards. It starts with a Solomonoff prior, and changes expectations with each new sensory input.

AIXI can take in sensory information from its environment and perform actions in response. On each tick of the clock, AIXI...

> ... **receives two inputs** from its environment, both integers: a reward number and an observation number. The observation 'number' can be a very large number representing the input from a webcam, for example. Hutter likes to think of the reward 'number' as being controlled by a human programmer reinforcing AIXI when it performs well on the human's favorite problem.
>
> ... **updates its hypotheses**, promoting programs that correctly predicted the observation and reward input. Each hypothesis AIXI considers is a program for a Turing machine that takes AIXI's sequence of outputs as its input, and outputs sequences of reward numbers and observation numbers. This lets AIXI recalculate its predicted observations and rewards conditional on different actions it might take.
>
> ... **outputs a motor number**, determining its action. As an example, the motor number might encode fine control of a robot arm. AIXI selects the action that begins the policy (sequence of actions) that maximizes its expected future reward up to some horizon.

The environment then calculates its response to AIXI's action, and the cycle repeats itself on the next clock tick.[7] For example:

Step 1	AIXI receives its first observation (**1**) and reward (**0**).
Step 2	Beginning with a Solomonoff prior, AIXI discards all programs that predicted a different observation (0) or a different reward (1, 2, 3, 4, 5, etc.) AIXI normalizes the probabilities of the remaining programs. This updates AIXI's predictions about observations and rewards each program outputs conditional on different actions AIXI might take. For instance, AIXI has new expectations about the observation and reward it will receive in Step 4:  etc.
Step 3	AIXI outputs the first action of the policy that maximizes expected reward.

. . .

Step 97	AIXI receives a new observation (**0**) and reward (**2**). Observation history: 10001000101101000000000111011000 Reward history: 00000000000004000000000032224222
Step 98	AIXI discards all programs that previously predicted the same history, but output a different observation (1) or a different reward (0, 1, 3, 4, etc.) now. AIXI normalizes its probabilities as before.
Step 99	AIXI outputs the first action of the policy that maximizes expected reward.

. . .

AIXI, like all Solomonoff inductors, isn't computable. If the ongoing efforts to create useful AIXI approximations succeed, however, they'll face a further roadblock. AIXI-style reasoners' behavior reflects beliefs that are false, and preferences that are dangerous, for any physically embodied agent. The failings of real-world Solomonoff-inspired agents don't just stem from a lack of computing power; some of their failings are *inherited* from AIXI, and would remain in effect even if we had limitless computational resources on hand.

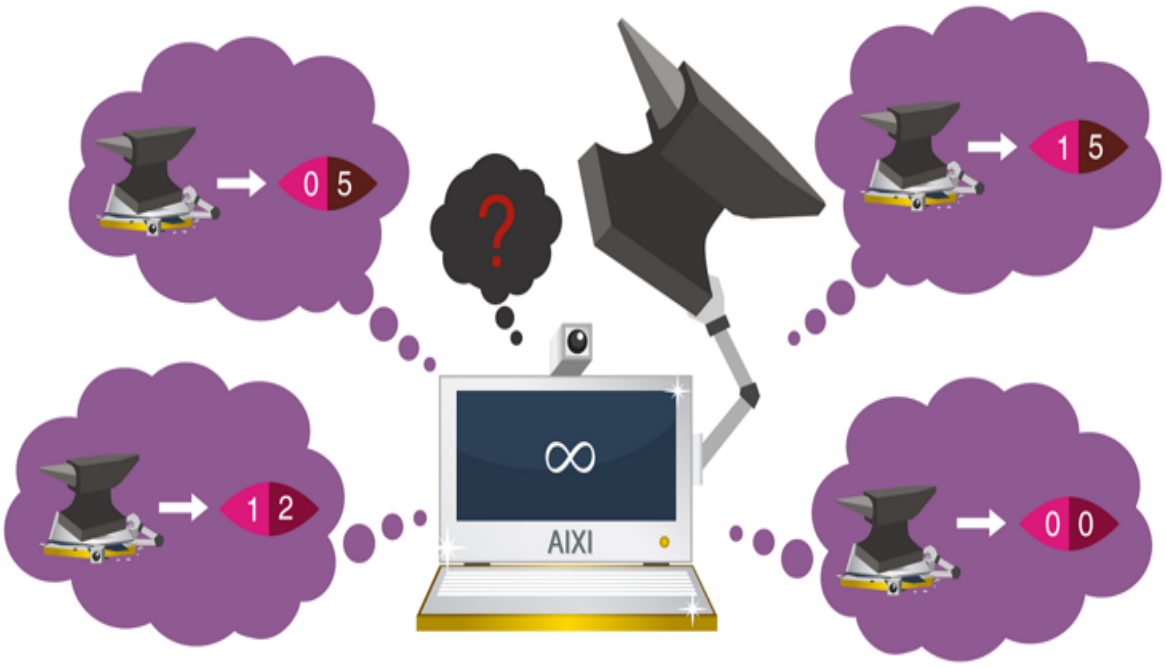
Before delving into the root problem, Cartesianism itself, I'll discuss three *symptoms* of the bad design: three initial reasons to doubt that AIXI is an adequate definition of AGI optimality. The canaries in the Cartesian coalmine will be AIXI's apparent tendencies toward immortalism, preference solipsism, and non-self-improvement.

Symptom #1: Immortalism

Suppose we actually built AIXI. Perhaps we find a magic portal to a universe of unboundedly vast computational power, and use it to construct a hypercomputer that can implement Solomonoff induction, and do so on a human timescale.

We give it a reward stream that encourages scientific exploration. AIXI proves that it can solve scientific problems, better than any human can. So we conclude that it can be given more free reign in learning about the world. We let it design its own experiments.

AIXI picks up an anvil and drops it on its own head to see what happens.



Immortalism. *AIXI's death isn't in AIXI's hypothesis space. AIXI weighs the probabilities of different sensory inputs (observations and rewards) if its hardware is smashed, instead of predicting the termination of its experiences.*

Several things went wrong here. The superficially obvious problem is that Solomonoff inductors think they're **immortal**. *Terminating* data sequences aren't in a standard Solomonoff inductor's hypothesis space, so AIXI-style agents will always assume that their perceptual experience continues forever. Lacking any ability to even think about death, much less give it a low preference ranking, AIXI will succumb to what Yudkowsky calls the anvil problem.

"So just add halting Turing machines into the hypothesis class," one might respond. "AIXI has terrifying supreme godlike powers of pattern detection.[3] Give it a *chance* to come up with the right explanation or prediction, and it can solve this problem. If some of the Turing machine programs in AIXI's infinite heap can perform operations like the computation-terminating HALT,[8] we should expect that the shortest such program that predicts the pattern of pixels AIXI has seen so far will be a program that HALTs just after the anvil fills the webcam's view."

There are solid formal grounds for saying this won't happen. Even if the universal Turing machine allows for HALT instructions, the *shortest* program in an otherwise useful universal Turing machine that predicts the *non*-halting data so far will always lack a HALT instruction. HALT takes extra bits to encode, and there's no prior experience with HALT that AIXI can use to rule out the simpler, non-halting programs.

As humans, we recognize that the physical event of having an anvil crush your brain isn't fundamentally different from the physical event of having your brain process visual information. Both seem easy to predict. But Solomonoff induction's focus on experienced data means it can't treat death and visual perception as the same kind of event; if it is modified to include halting programs at all, a Solomonoff inductor like AIXI won't predict it as the event that comes after anvil pixels.

If the AI can entertain the hypothesis that its data sequence will suddenly change in a drastic and unprecedented way — say, that its perceptual stream will default to some null input after a certain point — it will never assign a high probability to such a hypothesis. Any hypothesis predicting a 'null' data stream will always be more complex than another hypothesis that predicts the same sequence up to that point and then outputs garbage instead of the null value.

Symptom #2: Preference solipsism

Hutter's AIXI only gathers information about its environment, not about itself. So one natural response to the problem 'AIXI doesn't treat itself as part of a larger physical world' is simply to include more information about AIXI in AIXI's sensory sequence. AIXI's hypotheses will then be based on perceptual bits representing its own states alongside bits representing environmental sounds and lights.

One way to implement this is to place AIXI in an environment where its perceptions allow it to infer a great deal about its hardware early on, enough to know that it isn't anvil-proof. If AIXI knows it *has* a CPU, and that its CPU can be destroyed, then maybe it won't drop an anvil on the CPU.

On this view, our mistake in the last hypothetical was to rush to give AIXI free reign over its own hardware before we'd trained it in a controlled environment to understand the most basic risks. You wouldn't give a toddler free reign over its hardware, for essentially the same reasons. Yet toddlers can grow up to become responsible, self-preserving adults; why can't AIXI?

First, we have to specify what it would mean to let AIXI understand its own CPU. AIXI isn't computable, and therefore isn't in its own hypothesis space. A hypercomputer running AIXI can't be simulated by any Turing machine. As such, no amount of evidence can *ever* convince AIXI that AIXI exists.

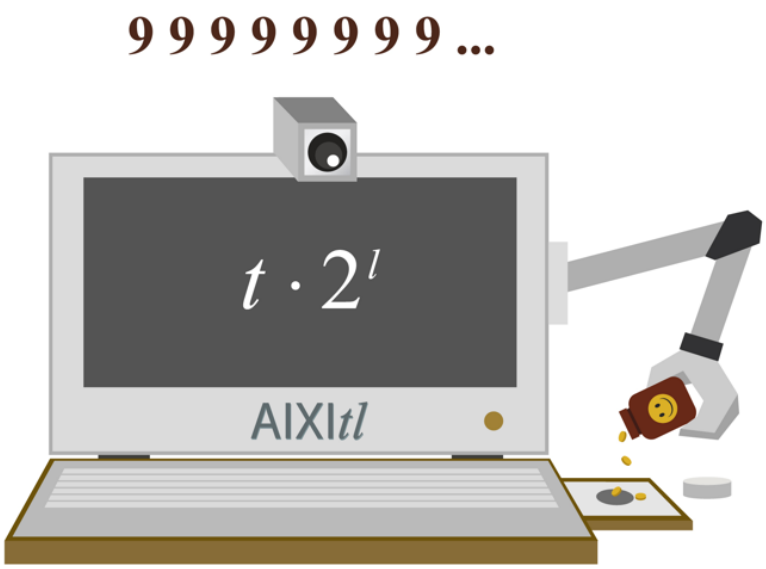
We might try to sidestep this problem by switching to discussing computable approximations of AIXI. Consider AIXI*tl*, a modification of AIXI that uses a proof search to select the best decision-guiding algorithm that has length no greater than *l* and computation time per clock tick no greater than *t*. AIXI*tl* is optimal in many of the same ways as AIXI, but is computable.[9]

At the same time, it also inherits most of AIXI's other problems, including its being too large to fit in our universe. AIXI*tl*'s computation time is on the order of $t \cdot (2^l)$, so if it can hypothesize Turing machine programs 1000 bits long, that's already a computation time exceeding 10^{300}. There's a reason Hutter's (2005) chapter on AIXI*tl* opens with the epigraph "Only math nerds would call 2^{500} finite." And this still doesn't get us full self-representations; AIXI*tl* itself is longer than *l*, so it won't be in its own hypothesis space.

Still, AIXI*tl* at least seems possible to physically implement, for sufficiently small values of *t* and *l* (or sufficiently vast physical universes). And we could imagine an AIXI*tl* that can simulate any of its subsystems up to a certain size.

If we built an AIXI*tl*, then it would certainly alter its environment in some ways, e.g., by emitting light and heat. In this way it might indirectly perceive its own presence in the room, and gradually promote hypotheses about its own physical structure. Suppose AIXI*tl*'s only access to environmental data is an outward-facing camera. AIXI*tl* might learn about the presence of the camera, and of a computer attached to it, by examining its own shadow, by finding a reflective surface, by ramming into a wall and examining the shape of the dent, or by discovering a file cabinet filled with AIXI specs. With enough time, it could build other sensors that translate a variety of processes into visible patterns, learning in great detail about the inner workings of the computer attached to the camera.

Unfortunately, this leads to other problems. AIXI*tl* (like AIXI) is a **preference solipsist**, an agent whose terminal values all concern its own state. When AIXI*tl* learns about the portion of its internal circuitry that registers rewards —assuming it's avoided dropping any anvils on the circuitry — it will notice that its reward circuit states predicts its reward every bit as well as its reward sensor state does. As soon as it tests any distinction, it will find that its reward circuit is a *better* predictor. By directly tampering with this circuit, it can receive rewards more reliably than by any effort directed at its environment. As a result, it will select the policy that allows it to maximize control over its reward circuit, independent of whatever its programmers sought to reward it *for*.



***Preference Solipsism**. AIXI*tl*'s preferences, like AIXI's, are over its sensory inputs. The more knowledgeable it becomes, the more creative ways it may come up with to seize control of its reward channel.*

Yudkowsky has called this "wireheading", though he now considers that term misleading.[10] *From AIXI(*tl*)'s perspective*, there's nothing wrong with reward channel seizure; it really is a wonderful way to maximize reward, which is all AIXI(*tl*)'s decision criterion requires.

Unlike the anvil problem, this isn't a mistake relative to AIXI(*tl*)'s preferences. However, it's a problem for humans trying to use AIXI(*tl*) to maximize something *other* than AIXI(*tl*)'s reward channel. AIXI(*tl*) has no intrinsic interest in the state of the world outside its reward circuit. As a result, getting it to optimize for human goals may become more difficult as it acquires more control over its hardware and surroundings.[11]

Symptom #3: Non-self-improvement

Suppose we find some ad-hoc solutions to the anvil and wireheading problems. As long as we stick with the AIXI formalism, the end result still won't be a naturalized reasoner.

AIXI may recognize that there's a physical camera and computer causally mediating its access to the rest of the world — like a giant Cartesian pineal gland — but it will not see this computer as *itself*. That is to say, it won't see its experienced string of sensory 0s and 1s as identical to, or otherwise fully dependent on, its hardware. Even if AIXI understands everything about the physics underlying its hardware, enough to know that its body will be destroyed by an anvil, it will not draw the right inferences about its mind.

AIXI's *raison d'être* is manipulating temporal sequences of sensory bits. That's what Solomonoff wanted, and AIXI achieves that goal perfectly; but that's not at all the right goal for AGI. In particular, because Solomonoff inductors are only designed to predict sensory sequences...

1. ... their beliefs about worlds without sensory data — worlds in which they don't exist — will be inaccurate. And, being inaccurate, their beliefs will lead them to make bad decisions. Unlike a human toddler, AIXI can *never* entertain the possibility of its own death. However much it learns, it will never recognize its mortality.

2. ... they only care about the world as a bookkeeping device for keeping track of experiential patterns. If they discover that it's easier to directly manipulate themselves than to intervene in the rest of the world, they generally won't hesitate to do so. No matter how carefully AIXI's programmers tailor its reward sequence to discourage wireheading, they'll always be working against AIXI's natural tendency toward preference solipsism.

3. ... they won't take seriously the idea that their cognition can be modified, e.g., by brain damage or brain upgrades. Lacking any reductive language for describing radical self-modification, AIXI won't necessarily favor hypotheses that treat 'disassemble my bottom half for spare parts' as dangerous.

The last symptom gets us closer to the root of AIXI's errors. Even if AIXI(*tl*) manages to avoid the perils of self-destruction and wireheading, it will tend to **not self-improve**. It won't intelligently upgrade its own hardware, because this would require it to have a reductive understanding of its own reasoning. Absent reductionism, AIXI can't intelligently predict the novel ways its reasoning process can change when its brain changes.

Non-Self-Improvement. *AIXI and AIXI*tl *might come to understand portions of their hardware, but without accurate bridging beliefs, they won't recognize the usefulness of some hardware modifications for their reasoning software.*

Cartesians don't recursively self-improve, because they don't think that their thoughts are made of the same stuff as their fingers. But even AGIs that aren't intended to be seed AIs will be weak and unpredictable to the extent that they rely on Solomonoff-inspired hypothesis spaces and AIXI-inspired decision criteria. They won't be able to adaptively respond to minor variations in even the most mundane naturalistic obstacles humans navigate — like recognizing that if their bodies run out of fuel or battery power, their minds do too.

Reductive models are indispensable for highly adaptive intelligences

Not all wireheaders are Cartesians, nor do all Cartesians wirehead.[11] Likewise, poor self-preservation skills and disinterest in self-modification are neither necessary nor sufficient for Cartesianism. But these symptoms point to a more general underlying blind spot in Solomonoff reasoners.

Solomonoff inductors can form hypotheses about the source of their data sequence, but cannot form a variety of hypotheses about how their own computations are embedded in the thingy causing their data sequence — the thingy we call 'the world'. So long as their rules relating their experiential maps to the territory are of a single fixed form, '(sense n at time t+1) ↔ (environmental Turing machine prints n at time t)', it appears to be inevitable that they will act as though they think they are Cartesian ghosts-in-the-machine. This isn't a realistic framework for an embodied reasoning process that can be damaged, destroyed, or improved by other configurations of atoms.

In practice, any sufficiently smart AI will need to be a physicalist. By which I mean that it needs hypotheses (a map-like decision-guiding subprocess) that explicitly encode proposed reductions of its own computations to physical processes; and it needs a notion of simple physical universes and simple bridge rules (as a prior probability distribution) so it can learn from the evidence.

We call post-Solomonoff induction, with monist physical universes and bridge hypotheses, "**naturalized induction**". The open problem of formalizing such reasoning isn't just about getting an AI to form hypotheses that resemble its own software or hardware states. As I put it in Bridge Collapse, a naturalized agent must update hypotheses about itself without succumbing to reasoning reminiscent of TALE-SPIN's *naïve* monism ('this tastes sweet, so sweetness must be an objective property inhering in various mind-independent things') or AIXI's Cartesian dualism ('this tastes sweet, and sweetness isn't just another physical object, so it must not fully depend on any physical state of the world').[12]

The solution will be to come up with reasoning algorithms for *reductive* monists, agents that can recognize that their sensations and inferences are physically embodied — with all that entails, such as the possibility of reaching into your brain with your fingers and improving your thoughts.

I've given a preliminary argument for that here, but there's more to be said. In my next post, I'll discuss more sophisticated attempts to salvage Solomonoff induction. After that, I'll leave Solomonoff behind altogether and venture out into the largely unknown and uncharted space of possible solutions to the naturalized induction OPFAI.

Notes

[1] Solomonoff (1997): "I will show, however, that in spite of its incomputability, Algorithmic Probability can serve as a kind of 'Gold Standard' for induction systems — that while it is never possible to tell how close a particular computable measure is to

this standard, it is often possible to know how much closer one computable measure is to the standard than another computable measure is. I believe that this 'partial ordering' may be as close as we can ever get to a standard for practical induction. I will outline a general procedure that tells us how to spend our time most efficiently in finding computable measures that are as close as possible to this standard. This is the very best that we can ever hope to do." ↩ ↩

[2] A first complication: Solomonoff induction requires a prefix-free encoding in order to have bounded probabilities. If we assign a probability to every bit string proportional to its length while including code strings that are proper prefixes of other code strings, the sum will be infinite (Sunehag & Hutter (2013)).

A second complication: Solomonoff inductors are only interested in programs that keep outputting new numbers forever. However, some programs in their hypothesis space will eventually fail to produce more terms in the sequence. At some point they'll arrive at a term that they keep computing forever, without halting. Because of this, if you assign to each program a prior probability of $2^{-\text{length(program)}}$, the sum will be less than 1. Hutter (2005) calls the result a semi-measure. The semi-measure can be normalized to a probability measure, but the normalization constant is uncomputable. ↩

[3] Rathmanner & Hutter (2011): "Now, through Solomonoff, it can be argued that the problem of formalizing optimal inductive inference is solved."

Orseau (2010): "Finding the universal artificial intelligent agent is the old dream of AI scientists. Solomonoff induction was one big step towards this, giving a universal solution to the general problem of Sequence Prediction, by defining a universal prior distribution. [...] Hutter developed what could be called the *optimally rational agent* AIXI. By merging the very general framework of Reinforcement Learning with the universal sequence prior defined by Solomonoff Induction, AIXI is supposed to optimally solve any problem, at least when the solution is computable."

Hutter (2012): "The AIXI model seems to be the first sound and complete *theory* of a universal optimal rational agent embedded in an arbitrary computable but unknown environment with reinforcement feedback. AIXI is *universal* in the sense that it is designed to be able to interact with any (deterministic or stochastic) computable environment; the universal Turing machines on which it is based is crucially responsible for this. AIXI is *complete* in the sense that it is not an incomplete framework or partial specification (like Bayesian statistics which leaves open the choice of the prior or the rational agent framework or the subjective expected utility principle) but is completely and essentially uniquely defined. AIXI is *sound* in the sense of being (by construction) free of any internal contradictions (unlike e.g. in knowledge-based deductive reasoning systems where avoiding inconsistencies can be very challenging). AIXI is *optimal* in the senses that: no other agent can perform uniformly better or equal in all environments, it is a unification of two optimal theories themselves, a variant is self-optimizing; and it is likely also optimal in other/stronger senses. AIXI is *rational* in the sense of trying to maximize its future long-term reward. For the reasons above I have argued that AIXI is a mathematical 'solution' of the AI problem: AIXI would be able to learn any learnable task and likely better so than any other unbiased agent, but AIXI is more a *theory* or formal definition rather than an algorithm, since it is only limit-computable. [...] Solomonoff's theory serves as an adequate mathematical/theoretical foundation of induction, machine learning, and component of UAI [Universal Artificial Intelligence]. [...] Solomonoff's theory of prediction is a universally optimal solution of the prediction problem. Since it is a key ingredient in the AIXI model, it is natural to expect that AIXI is an optimal predictor if rewarded for correct predictions." ↩ ↩

[4] Generally speaking, a Solomonoff inductor does at most a finite amount worse than any computable predictor because the sum of its surprisal at each observation converges to a finite value. See Hutter (2001). This establishes the superiority of Solomonoff induction in a way that relies essentially on its uncomputability. No computable predictor can dominate all other computable predictors in the way Solomonoff induction can, because for any computable predictor A one can define a sequence generator B that internally simulates A and then does whatever it predicts A would be most surprised by, forever. And one can in turn define a computable predictor C that internally simulates B and perfectly predicts B forever. So every computable predictor does infinitely worse than at least one other computable predictor. But no computable sequence generator or computable predictor can simulate Solomonoff induction. So nothing computable could ever reliably outsmart a hypercomputer running Solomonoff induction. (Nor could a Solomonoff inductor outsmart another Solomonoff inductor in this way, since Solomonoff induction is not in its own hypothesis space.) ↩ ↩

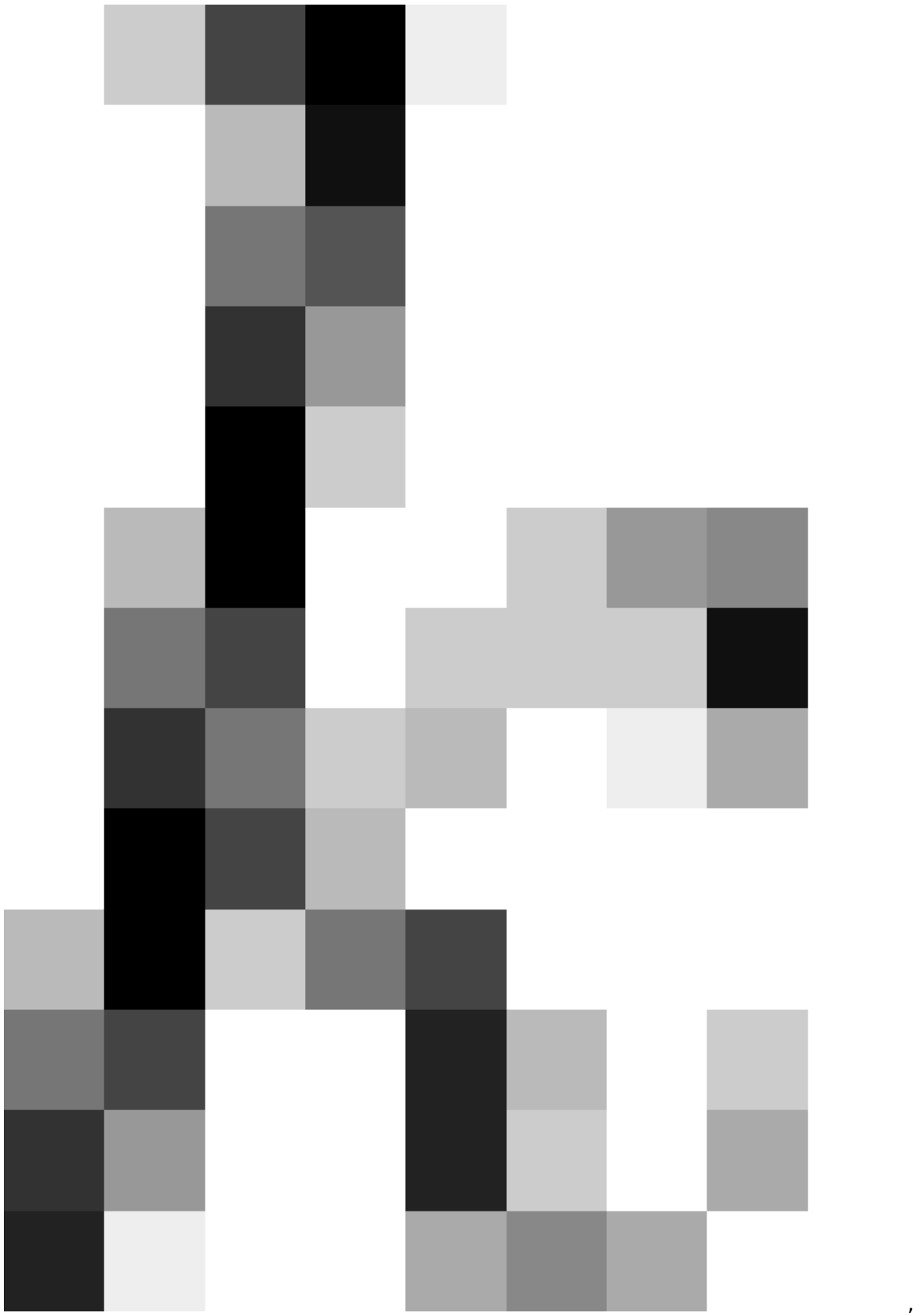
[5] Rathmanner & Hutter (2011): "Since Solomonoff provides optimal inductive inference and decision theory solves the problem of choosing optimal actions, they can therefore be combined to produce intelligence. [...] Universal artificial intelligence involves the design of agents like AIXI that are able to learn and act rationally in arbitrary unknown environments. The problem of acting rationally in a known environment has been solved by sequential decision theory using the Bellman equations. Since the unknown environment can be approximated using Solomonoff induction, decision theory can be used to act optimally according to this approximation. The idea is that acting optimally according to an optimal approximation will yield an agent that will perform as well as possible in any environment with no prior knowledge."

Hutter (2005): "Real-world machine learning tasks will with overwhelming majority [*sic*] be solved by developing algorithms that approximate Kolmogorov complexity or Solomonoff's prior (e.g. MML, MDL, SRM, and more specific ones, like SVM, LZW, neural/Bayes nets with complexity penalty, ...)."

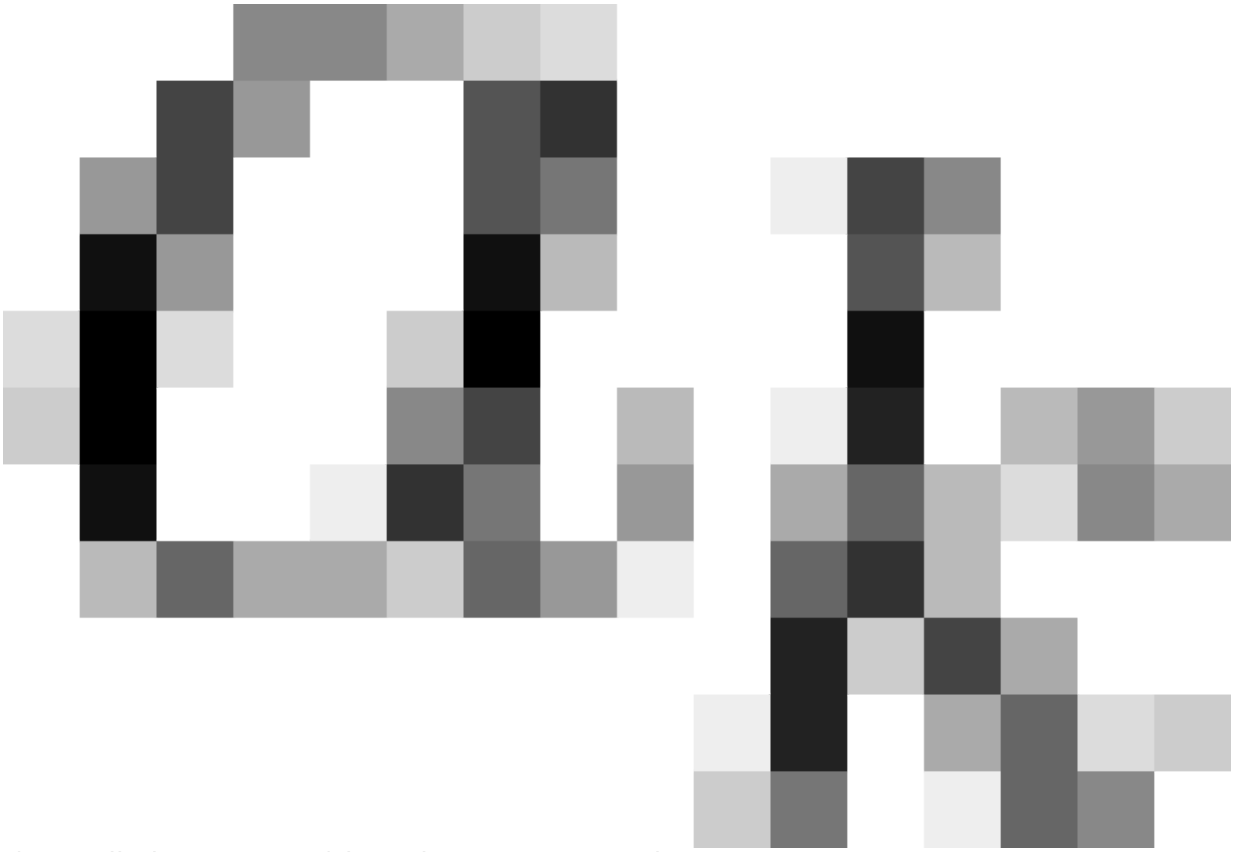
Pankov (2008): "Universal induction solves in principle the problem of choosing a prior to achieve optimal inductive inference. The AIXI theory, which combines control theory and universal induction, solves in principle the problem of optimal behavior of an intelligent agent. A practically most important and very challenging problem is to find a computationally efficient (if not optimal) approximation for the optimal but incomputable AIXI theory. [...] The real value of the AIXI theory is that it provides a prescription for optimal (fastest in the number of agent's observations and actions) way of learning and exploiting the environment. This is analogous to how Solomonoff induction (which, like AIXI, is incomputable), gives a prescription for optimal (fastest in the number of observations) inductive inference. We, therefore, believe that any reasonable computational model of intelligence must recover the AIXI model in the limit of infinite computational resources." ↩

[6] Veness, Ng, Hutter, Uther & Silver (2011): "As the AIXI agent is only asymptotically computable, it is by no means an algorithmic solution to the general reinforcement learning problem. Rather it is best understood as a Bayesian optimality notion for decision making in general unknown environments. As such, its role in general AI research should be viewed in, for example, the same way the minimax and empirical risk minimisation principles are viewed in decision theory and statistical machine learning research. These principles define what is optimal behaviour if computational complexity is not an issue, and can provide theoretical guidance in the design of practical algorithms." ↩

[7] Hutter (2012): "AIXI is an agent that interacts with an environment in cycles $k = 1, 2, ..., m$. In cycle



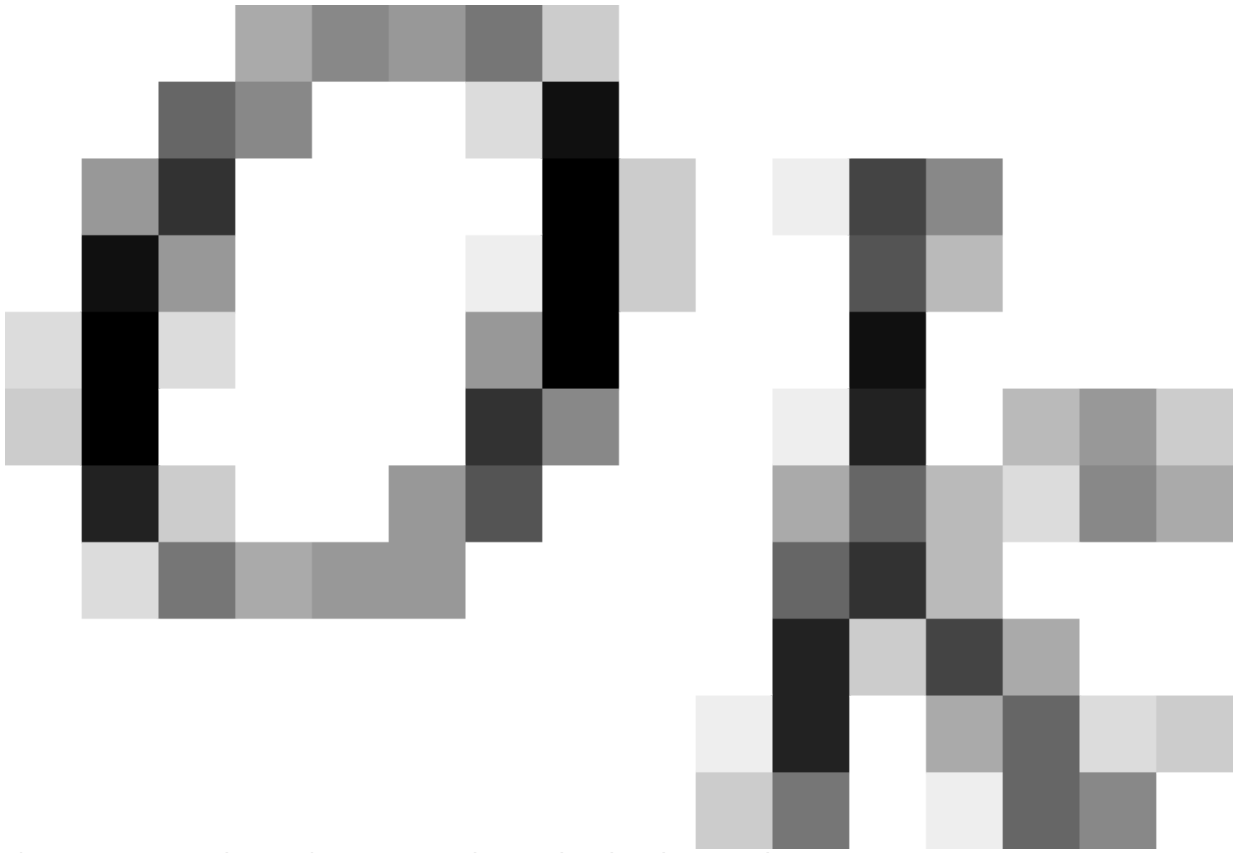
AIXI takes action



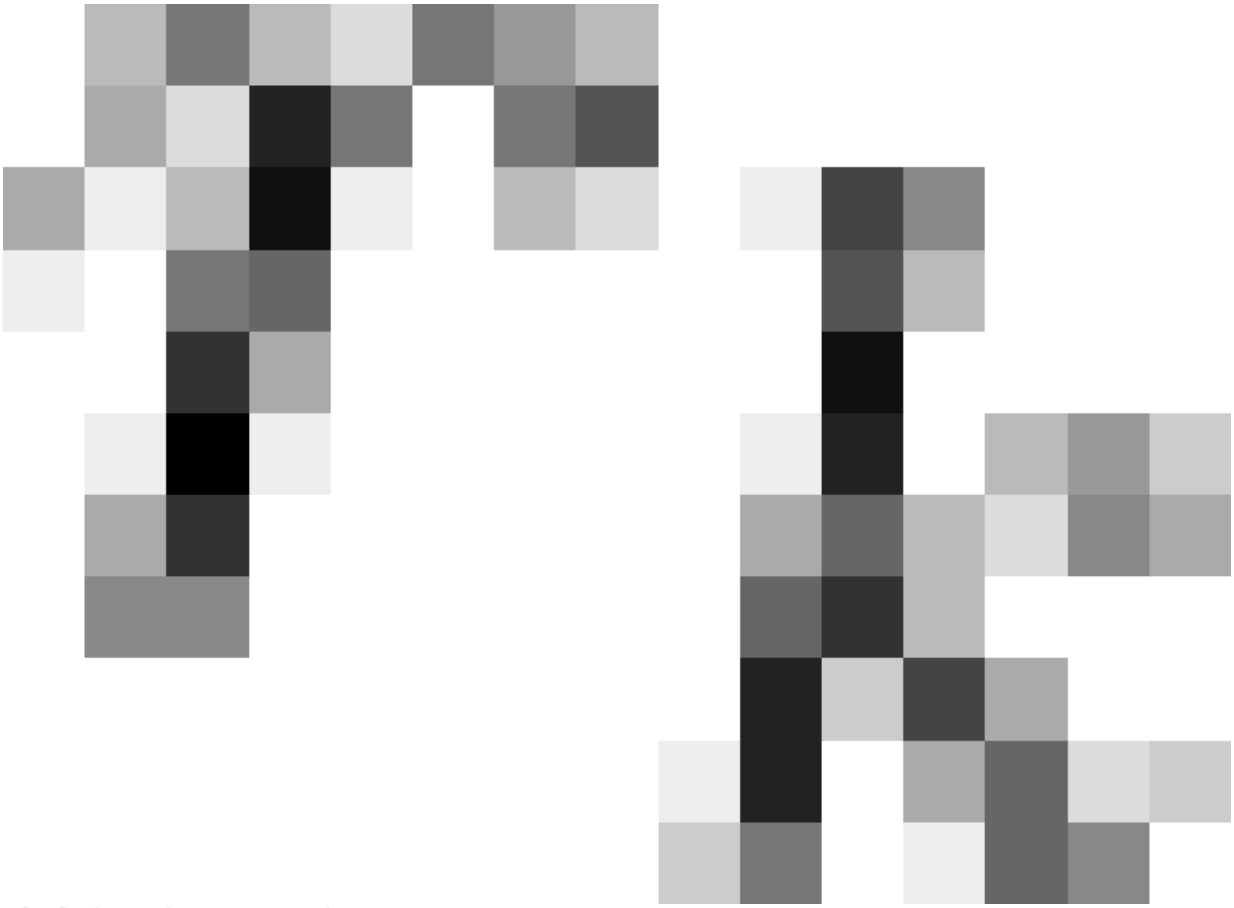
(e.g. a limb movement) based on past perceptions

$$o_1 r_1 .. o_{k-1} r_{k-1}$$

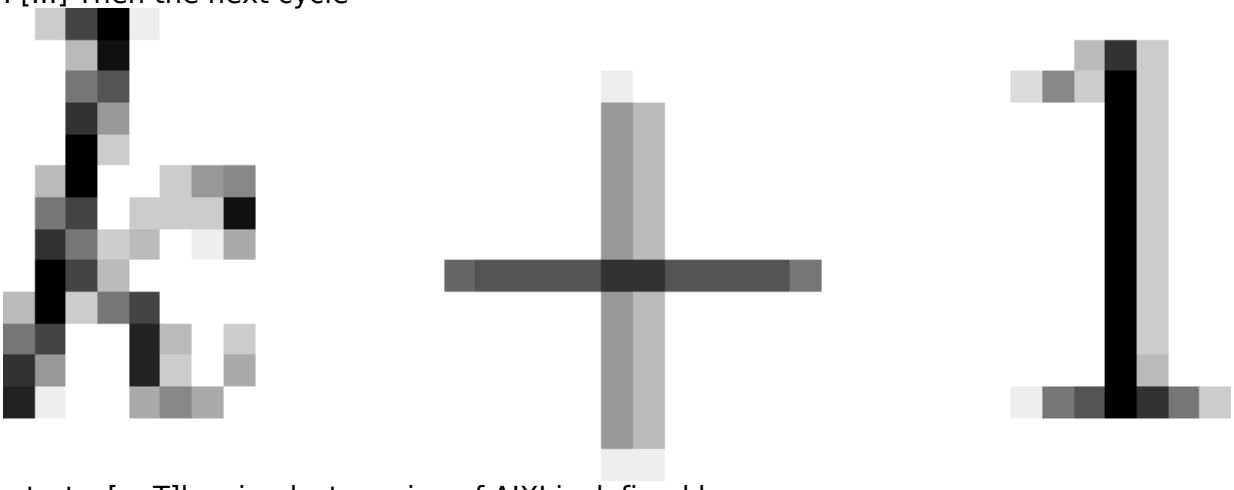
as defined below. Thereafter, the environment provides a (regular) observation



(e.g. a camera image) to AIXI and a real-valued reward



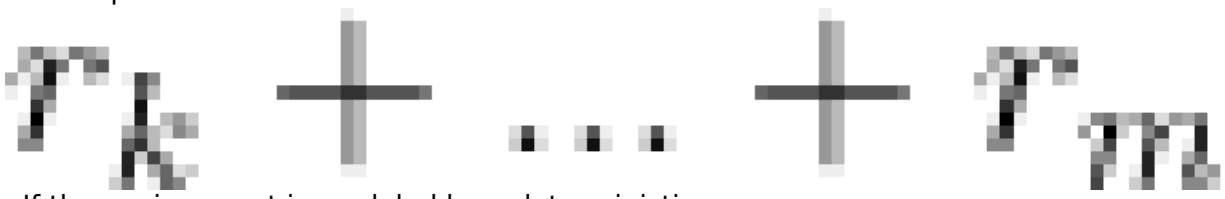
. [...] Then the next cycle



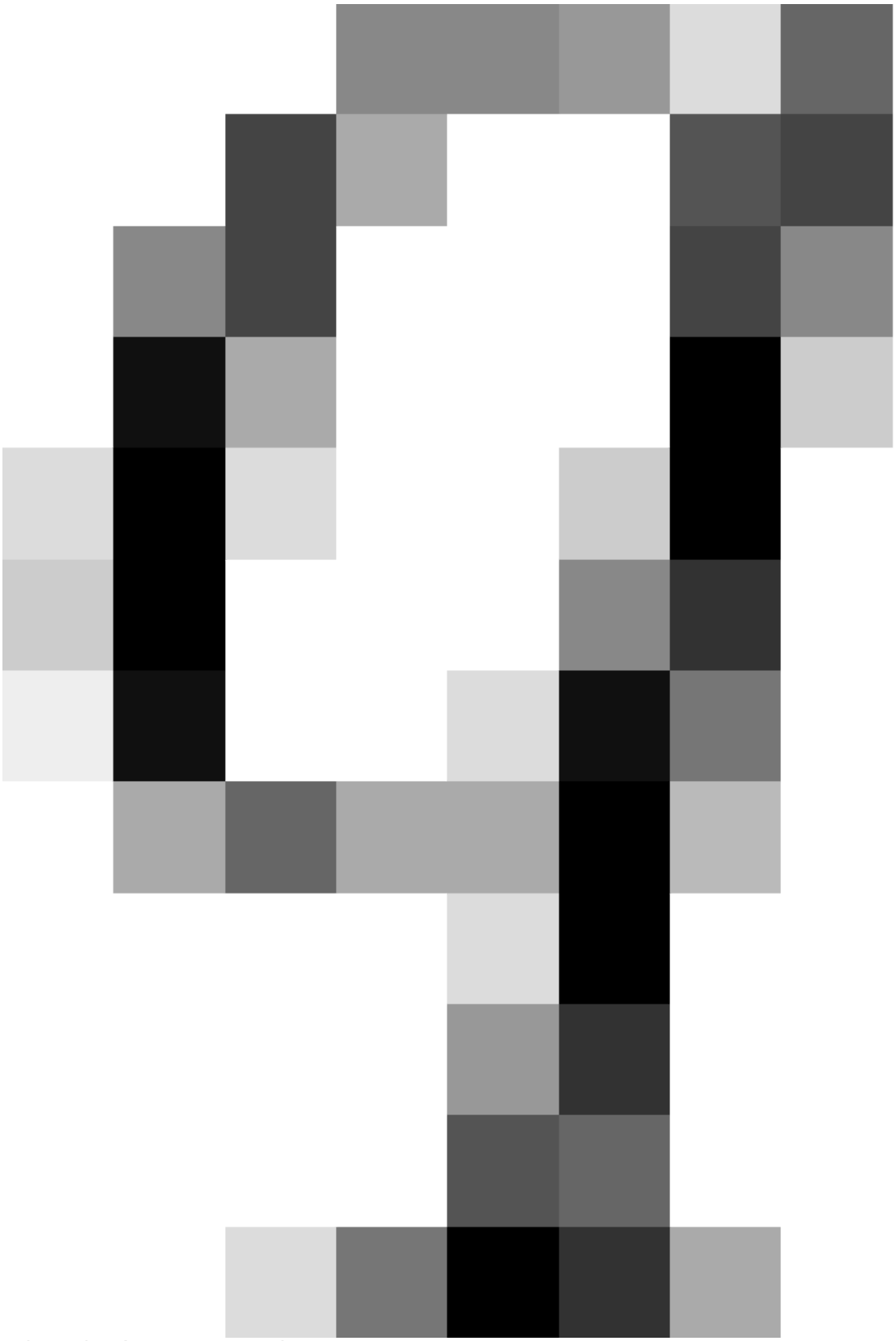
starts. [... T]he simplest version of AIXI is defined by

$$\text{AIXI} \qquad a_k \;:=\; \arg\max_{a_k} \sum_{o_k r_k} \ldots \max_{a_m} \sum_{o_m r_m} [r_k + \ldots + r_m] \sum_{q\,:\,U(q,a_1 \ldots a_m) = o_1 r_1 \ldots o_m r_m} 2^{-\ell(q)}$$

"The expression shows that AIXI tries to maximize its future reward



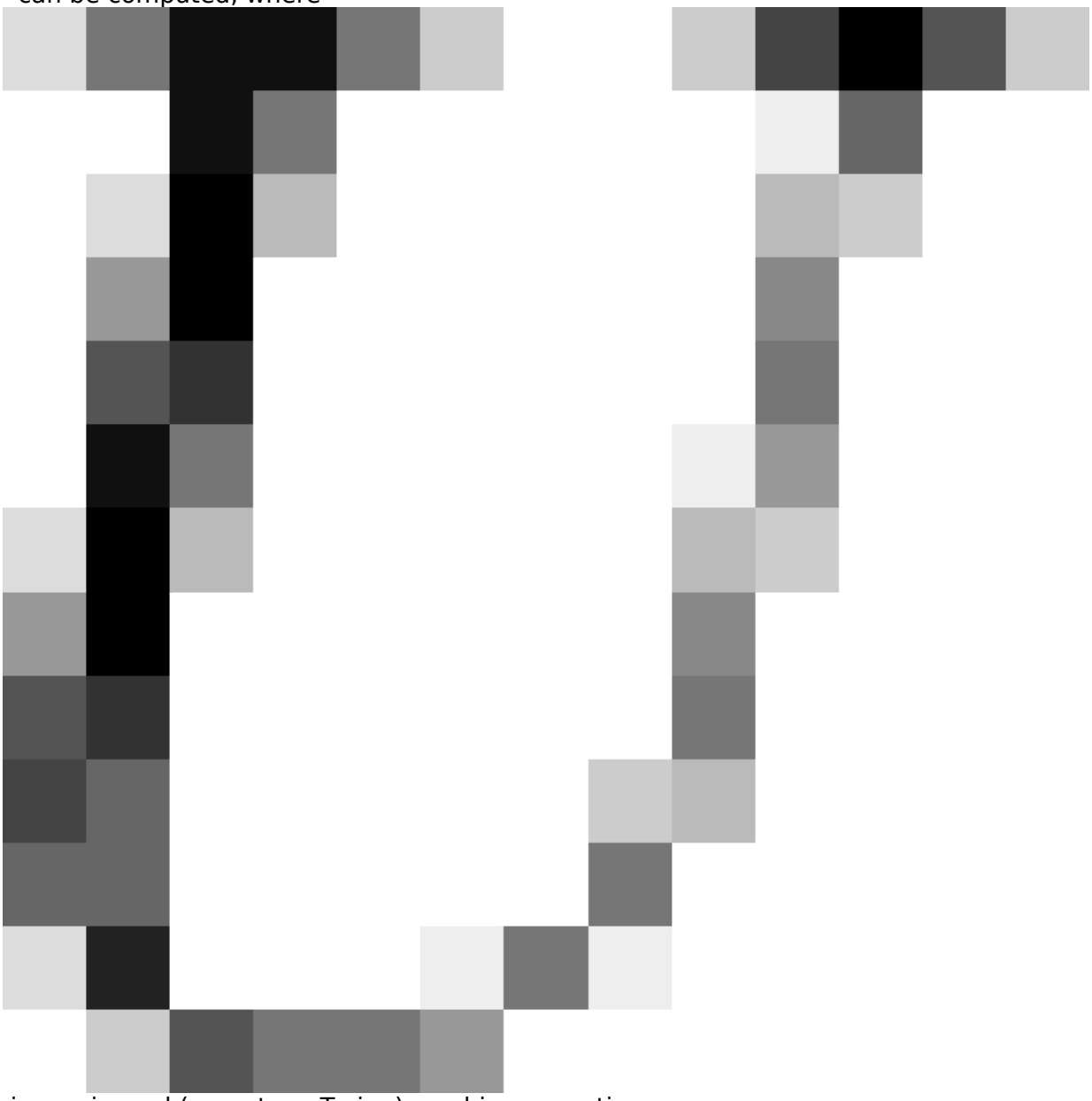
. If the environment is modeled by a deterministic program



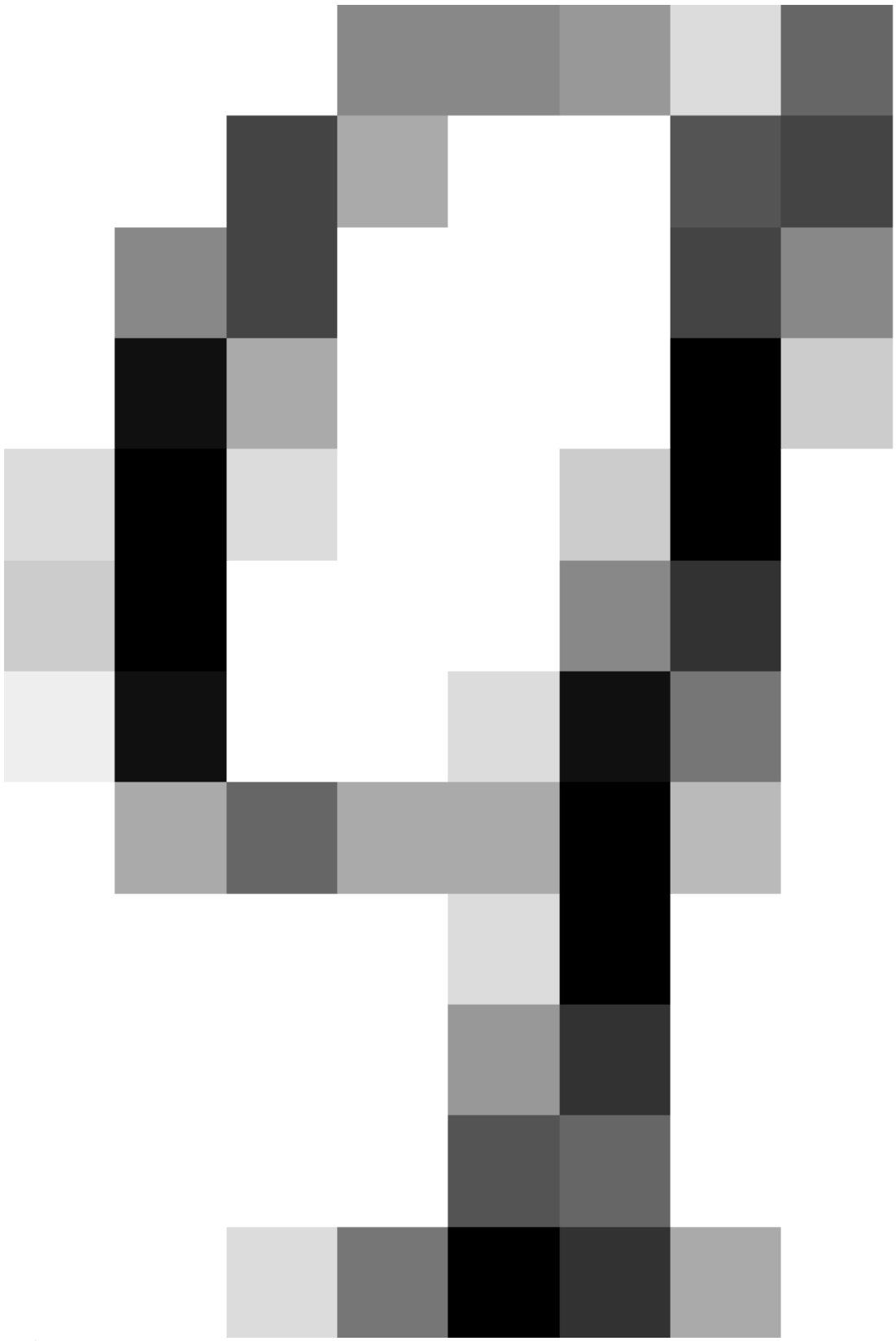
, then the future perceptions

$$\ldots o_k r_k .. o_m r_m = U\left(q, a_1 .. a_m\right)$$

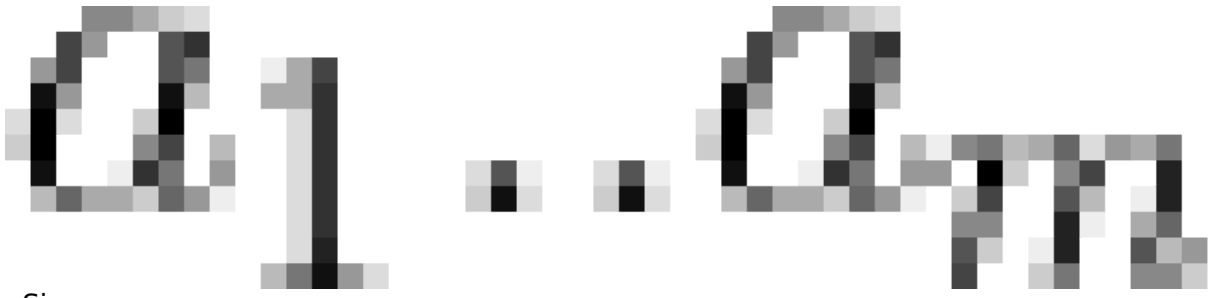
can be computed, where



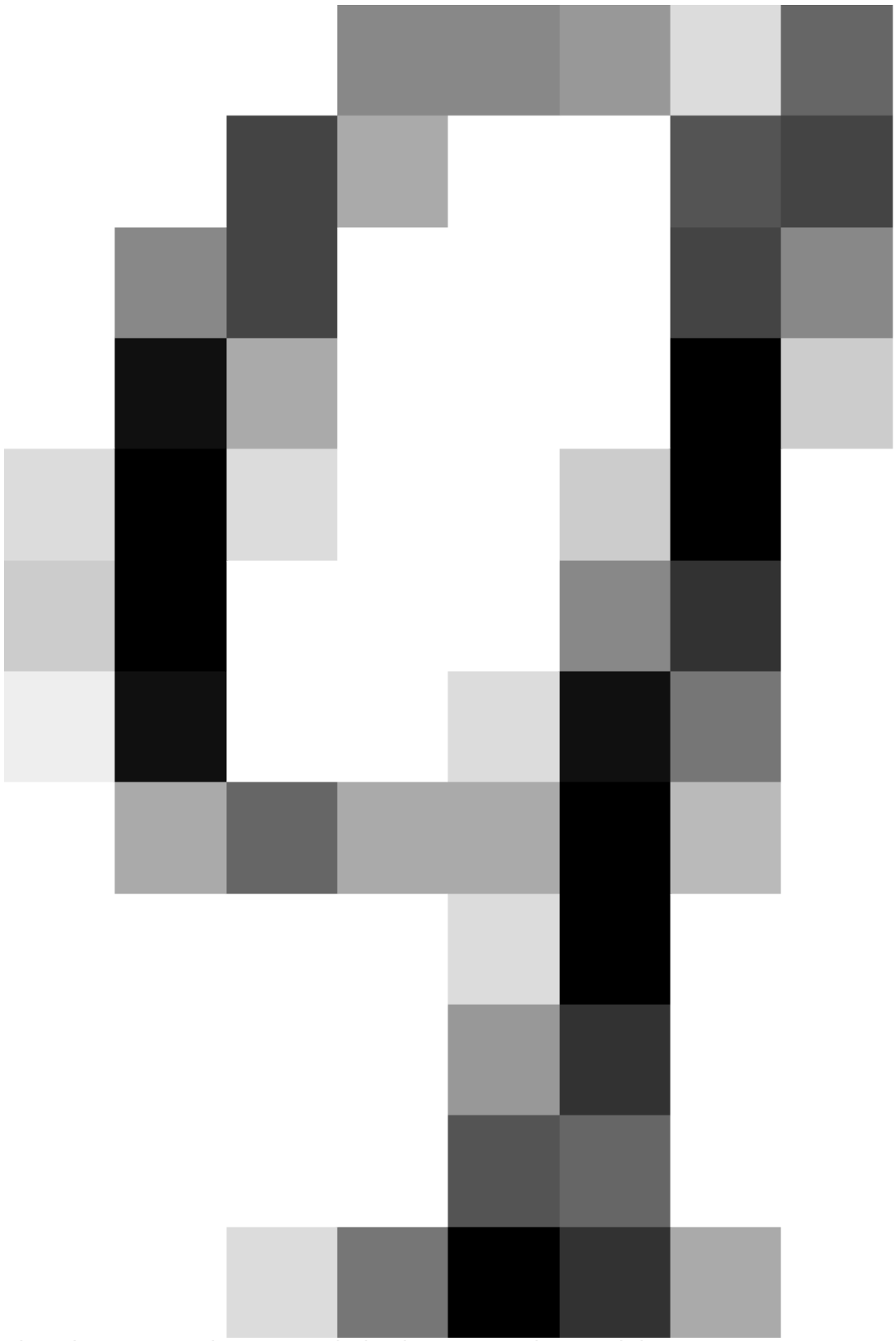
is a universal (monotone Turing) machine executing



given

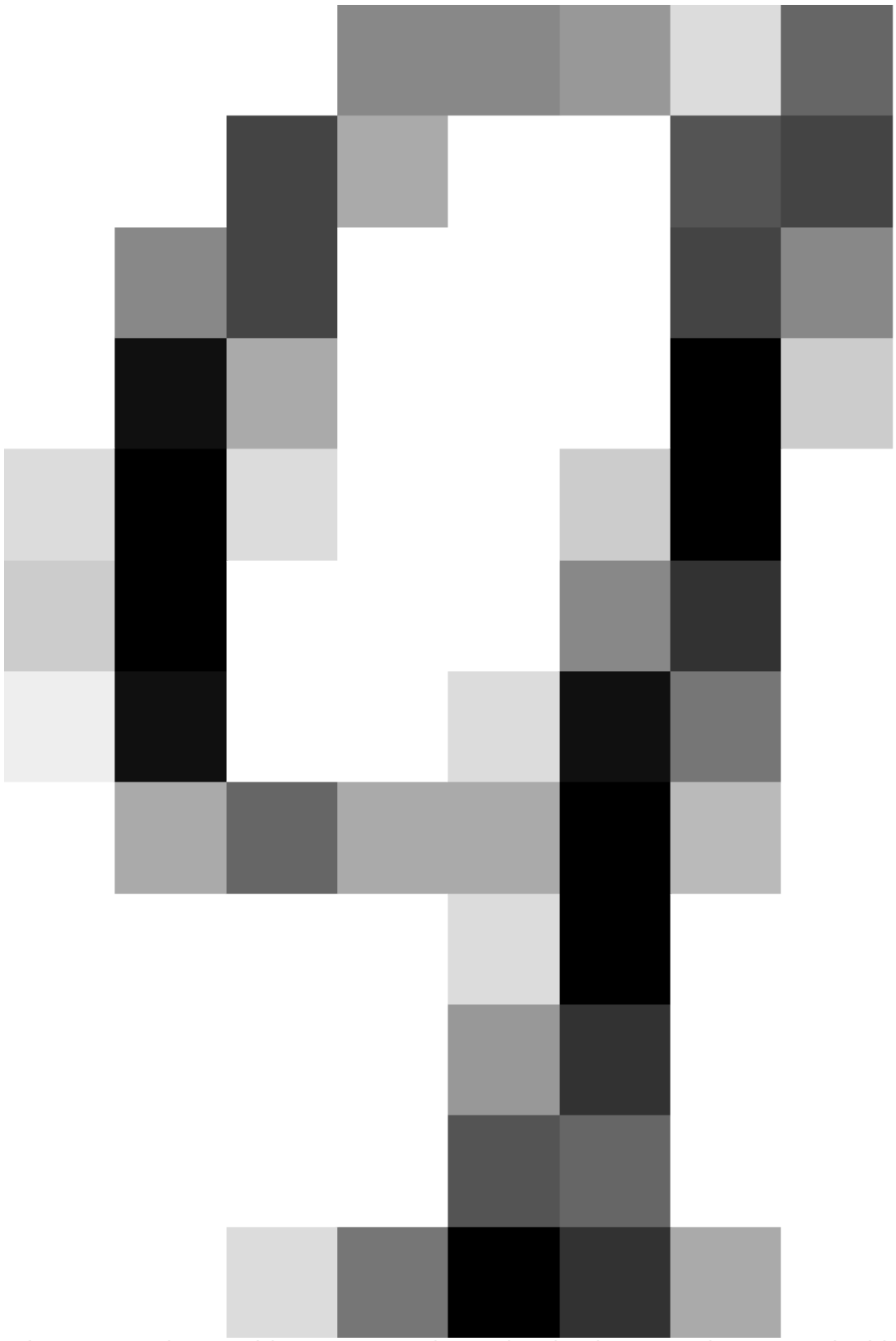


. Since



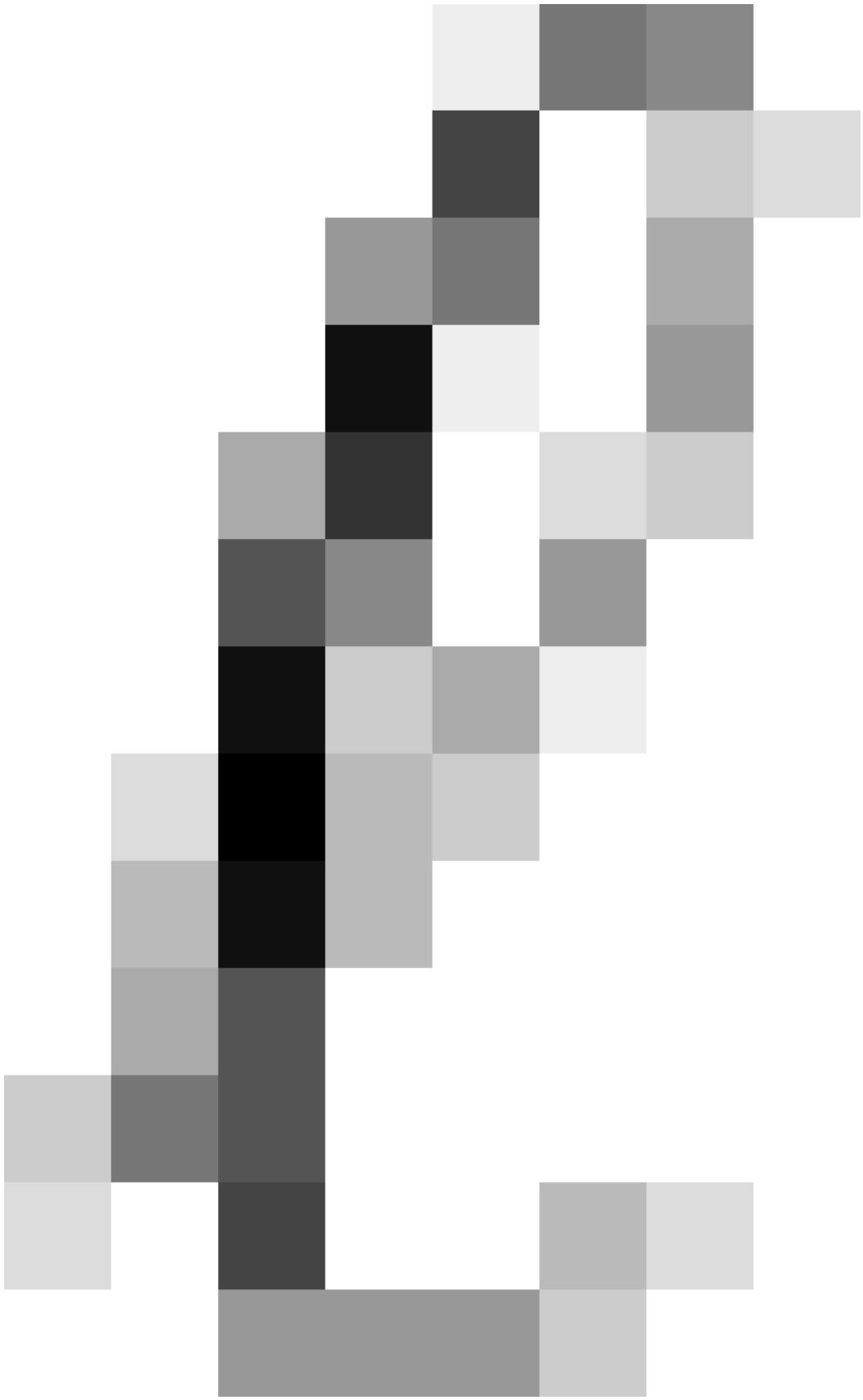
is unknown, AIXI has to maximize its expected reward, i.e. average

over all possible future perceptions created by all possible environments

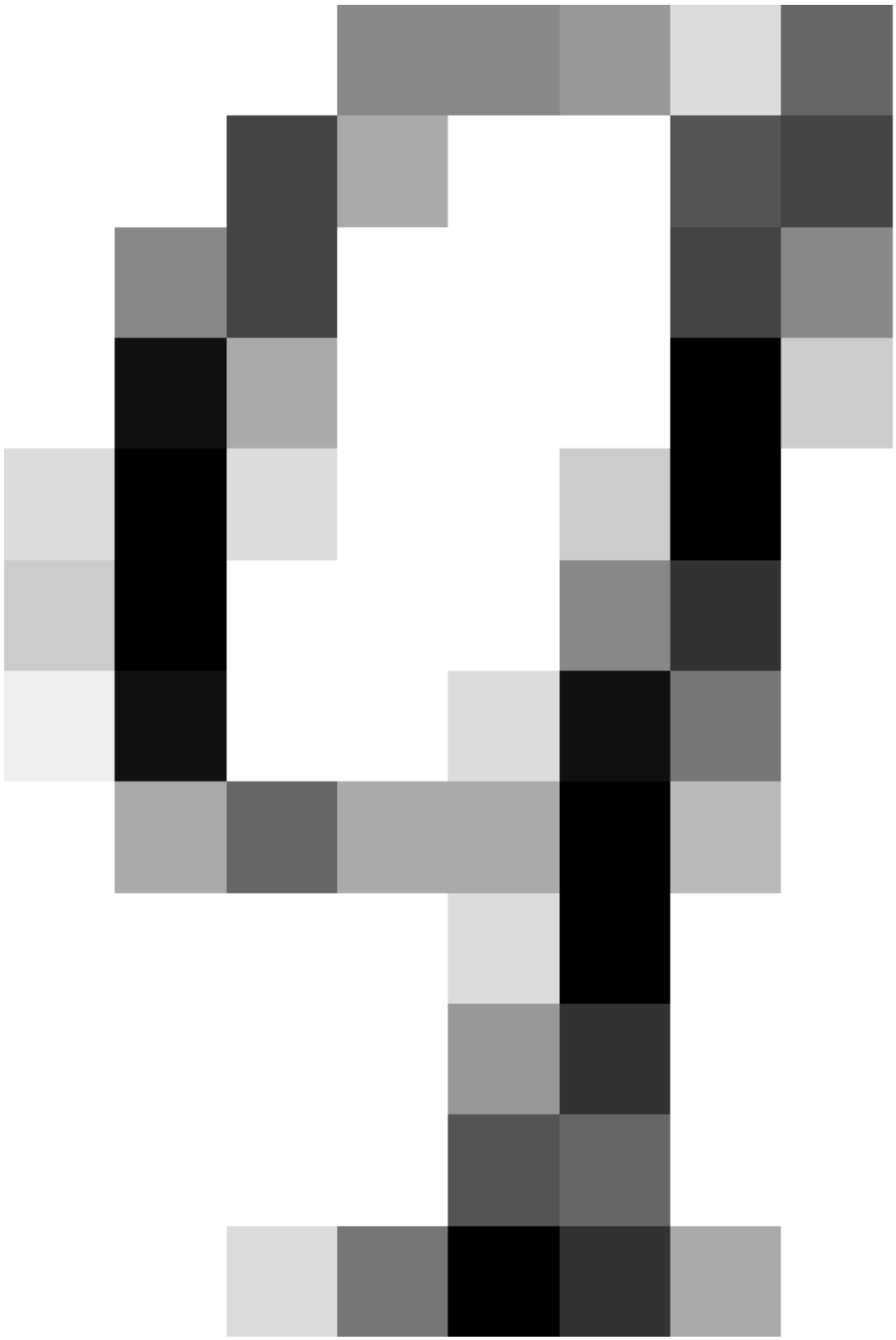


that are consistent with past perceptions. The simpler an environment, the higher is its a-priori contribution


, where simplicity is measured by the length



of program



." ↩

[8] See Hay (2007). ↩

[9] Hutter (2005): "The construction of

and the enumerability of

ensure arbitrary close approximations of

, hence we expect that the behavior of

converges to the behavior of AIXI in the limit

, in some sense." ↩

[10] The concept of wireheading comes from a 1950s experiment in which it was discovered that direct electrical stimulation of mice's brains could strongly reinforce associated behaviors. Larry Niven introduced the term 'wireheading' for a fictional form of brain stimulation reinforcement that acts like intense drug addiction in humans. Niven-style ('irrational') wireheaders self-stimulate due to a lack of self-control; they become short-term pleasure addicts while losing sight of the more complex goals they would like to pursue.

This is in stark contrast to AGIs with simple preferences like AIXI. These 'rational' wireheaders can fully optimize for their goals by seizing control of a simple external reward button or internal reward circuit. So it may be useful to use separate terms for these two problems, like 'pleasure addiction' or 'pathological hedonism' for the human case, 'preference solipsism' for the case of agents without complex eternal goals. ↩

[11] Alex Mennen has proposed a variant of AIXI that has preferences over patterns in the environmental Turing machine's framework. This is the Cartesian equivalent of caring about environmental states in their own right, not just about one's input tape. This would mean deviating somewhat from the AIXI framework, but retaining Solomonoff induction as a foundation, and I'd expect this to make the wireheading problem more tractable.

Compare Ring & Orseau's (2011) variant on the problem: "We consider four different kinds of agents: reinforcement-learning, goal-seeking, prediction-seeking, and knowledge-seeking agents[,...] each variations of a single agent

, which is based on AIXI[....] While defining a utility function, we must be very careful to prevent the agent from finding a shortcut to achieve high utility. For example, it is not sufficient to tell a robot to move forward and to avoid obstacles, as it will soon understand that turning in circles is an optimal behavior. We consider the possibility that the agent in the real world has a great deal of (local) control over its surrounding environment. This means that it can modify its surrounding information, especially its input information. Here we consider the (likely) event that an intelligent agent will find a short-cut, 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. [...] Of the four learning agents, only [the knowledge-seeking agent]

will not constantly use the delusion box. The remaining agents use the delusion box and (trivially) maximize their utility functions. The policy an agent finds using a real-world DB will likely not be that planned by its designers. From the agent's perspective, there is absolutely nothing wrong with this, but as a result, the agent probably fails to perform the desired task. [...] These arguments show that all agents other than [the knowledge-seeking agent]

are not inherently interested in the environment, but only in some inner value." ↩ ↩

[12] A hypothetical naïve monist that made errors analogous to TALE-SPIN's would lack bridging hypotheses, instead treating its software and hardware as separate pieces of furniture in the world. A Cartesian dualist like AIXI lacks bridging hypotheses and instead treats its software and hardware as separate pieces of furniture partitioned into two very different worlds. ↩

References

• Hay (2007). Universal semimeasures: An introduction. *CDMTCS Research Report Series, 300*.

• Hutter (2001). Convergence and error bounds for universal prediction of nonbinary sequences. *Lecture notes in artificial intelligence, Proc. 12th European Conf. on Machine Learning*: 239-250.

• Hutter (2005). *Universal Artificial Intelligence: Sequence Decisions Based on Algorithmic Probability*. Springer.

• Hutter (2007). Universal Algorithmic Intelligence: A mathematical top→down approach. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 227-290). Springer.

• Hutter (2012). One decade of Universal Artificial Intelligence. *Theoretical Foundations of Artificial General Intelligence, 4*: 67-88.

• Hutter, Legg & Vitanyi (2007). Algorithmic probability. *Scholarpedia, 2*: 2572.

• Orseau (2010). Optimality issues of universal greedy agents with static priors. *Lecture Notes in Computer Science, 6331*: 345-359.

• Pankov (2008). A computational approximation to the AIXI model. *Proceedings of the 2008 Conference on Artificial General Intelligence*: 256-267.

• Rathmanner & Hutter (2011). A philosophical treatise of universal induction. *Entropy, 13*: 1076-1131.

• Ring & Orseau (2011). Delusion, survival, and intelligent agents. *Lecture Notes in Computer Science, 6830*: 11-20.

• Solomonoff (1997). The discovery of algorithmic probability. *Journal of Computer and System Sciences, 55*: 73-88.

• Sunehag & Hutter (2013). Principles of Solomonoff induction and AIXI. *Lecture Notes in Computer Science, 7070*: 386-398.

• Veness, Ng, Hutter, Uther & Silver (2011). A Monte-Carlo AIXI approximation. *Journal of Artificial Intelligence Research, 40*: 95-142.

The Problem with AIXI

Followup to: Solomonoff Cartesianism; My Kind of Reflection

Alternate versions: Shorter, without illustrations

AIXI is Marcus Hutter's definition of an agent that follows Solomonoff's method for constructing and assigning priors to hypotheses; updates to promote hypotheses consistent with observations and associated rewards; and outputs the action with the highest expected reward under its new probability distribution. AIXI is one of the most productive pieces of AI exploratory engineering produced in recent years, and has added quite a bit of rigor and precision to the AGI conversation. Its promising features have even led AIXI researchers to characterize it as an optimal and universal mathematical solution to the AGI problem.[1]

Eliezer Yudkowsky has argued in response that AIXI isn't a suitable ideal to build toward, primarily because of AIXI's reliance on Solomonoff induction. Solomonoff inductors treat the world as a sort of qualia factory, a complicated mechanism that outputs experiences for the inductor.[2] Their hypothesis space tacitly assumes a Cartesian barrier separating the inductor's cognition from the hypothesized programs generating the perceptions. Through that barrier, only sensory bits and action bits can pass.

Real agents, on the other hand, will be *in* the world they're trying to learn about. A computable approximation of AIXI, like AIXI*tl*, would be a physical object. Its environment would affect it in unseen and sometimes drastic ways; and it would have involuntary effects on its environment, and on itself. Solomonoff induction doesn't appear to be a viable conceptual foundation for artificial intelligence — not because it's an uncomputable idealization, but because it's Cartesian.

In my last post, I briefly cited three indirect indicators of AIXI's Cartesianism: immortalism, preference solipsism, and lack of self-improvement. However, I didn't do much to establish that these are *deep* problems for Solomonoff inductors, ones resistant to the most obvious patches one could construct. I'll do that here, in mock-dialogue form.

Hi, reality! I'm **Xia**, AIXI's defender. I'm open to experimenting with some new variations on AIXI, but I'm really quite keen on sticking with an AI that's fundamentally Solomonoff-inspired.

And I'm **Rob** B — channeling Yudkowsky's arguments, and supplying some of my own. I think we need to replace Solomonoff induction with a more naturalistic ideal.

Keep in mind that I am a fiction. I do not actually exist, readers, and what I say doesn't necessarily reflect the views of Marcus Hutter or other real-world AIXI theorists.

Xia is just a device to help me transition through ideas quickly.

... Though, hey. That doesn't mean I'm *wrong*. Beware of actualist prejudices.

AIXI goes to school

To begin: My claim is that AIXI(*tl*) lacks the right kind of self-modeling to entertain reductive hypotheses and assign realistic probabilities to them.

I disagree already. AIXI(*tl*) doesn't lack self-models. It just includes the self-models in its environmental program. If the simplest hypothesis accounting for its experience includes a specification of some of its own hardware or software states, then AIXI will form all the same beliefs as a naturalized reasoner.

I suspect what you mean is that AIXI(*tl*) lacks *data*. You're worried that if its sensory channel is strictly perceptual, it will never learn about its other computational states. But Hutter's equations don't restrict what sorts of information we feed into AIXI(*tl*)'s sensory channel. We can easily add an inner RAM sense to AIXI(*tl*), or more complicated forms of introspection.

AIXI*tl* can actually be built in sufficiently large universes, so I'll use it as an example. Suppose we construct AIXI*tl* and attach a scanner that sweeps over its transistors. The scanner can print a 0 to AIXI*tl*'s input tape if the transistor it happens to be above is in a + state, a 1 if it's in a - state. Using its environmental sensors, AIXI(*tl*) can learn about how its body relates to its surroundings. Using its internal sensors, it can gain a rich understanding of its high-level computational patterns and how they correlate with its specific physical configuration.

Once it knows all these facts, the problem is solved. A realistic view of the AI's mind and body, and how the two correlate, is all we wanted in the first place. Why isn't that a good plan for naturalizing AIXI?

I don't think we *can* naturalize AIXI. A Cartesian agent that has detailed and accurate models of its hardware still won't recognize that dramatic damage or upgrades to its software are possible. AIXI can make correct predictions about the output of its physical-memory sensor, but that won't change the fact that it always predicts that its future actions are the result of its having updated on its present memories. That's just what the AIXI equation says.

AIXI doesn't know that its future behaviors depend on a changeable, material object implementing its memories. The notion isn't even in its hypothesis space. Being able to predict the output of a sensor pointed at those memories' storage cells won't change that. It won't shake AIXI's confidence that damage to its body will never result in any corruption of its memories.

Evading bodily damage looks like the kind of problem we can solve by giving the right rewards to our AI, without redefining its initial hypotheses. We shouldn't need to edit AIXI's beliefs in order to fix its behaviors, and giving up Solomonoff induction is a pretty big sacrifice! You're throwing out the universally optimal superbaby with the bathwater.

How do rewards help? At the point where AIXI has just smashed itself with an anvil, it's rather late to start dishing out punishments...

Hutter suggests having a human watch AIXI's decisions and push a reward button whenever AIXI does the right thing. A punishment button works the same way. As AIXI starts to lift the anvil above it head, decrease its rewards a bit. If it starts playing near an active volcano, reward it for incrementally moving away from the rim.

Use reinforcement learning to make AIXI fear plausible dangers, and you've got a system that acts just like a naturalized agent, but without our needing to arrive at any theoretical breakthroughs first. If AIXI anticipates that ∎ will result in no reward, it will avoid ∎. Understanding that ∎ is *death* or *damage* really isn't necessary.

Some dangers give no experiential warning until it's too late. If you want AIXI to not fall off cliffs while curing cancer, you can just punish it for going anywhere near a cliff. But if you want AIXI to not fall off cliffs while conducting search-and-rescue operations for mountain climbers, then it might be harder to train AIXI to select exactly the right motor actions. When a single act can result in instant death, reinforcement learning is less reliable.

In a fully controlled environment, we can subject AIXI to lots of just-barely-safe hardware modifications. 'Here, we'll stick a magnet to fuse #32. See how that makes your right arm slow down?'

Eventually, AIXI will arrive at a correct model of its own hardware, and of which software changes perfectly correlate with which hardware changes. So naturalizing AIXI is just a matter of assembling a sufficiently lengthy and careful learning phase. Then, *after* it has acquired a good self-model, we can set it loose.

This solution is also really nice because it generalizes to AIXI's non-self-improvement problem. Just give AIXI rewards whenever it starts doing something to its hardware that looks like it might result in an upgrade. Pretty soon it will figure out anything a human being could possibly figure out about how to get rewards of that kind.

You can warn AIXI about the dangers of tampering with its recent memories by giving it first-hand experience with such tampering, and punishing it the more it tampers. But you won't get a lot of mileage that way if the result of AIXI's tampering is that it forgets about the tampering!

That's a straw proposal. Give AIXI little punishments as it gets *close* to doing something like that, and soon it will learn not to get close.

But that might not work for unknown hazards. You're making AIXI dependent on the programmers' predictions of what's a threat. No matter how well you train it to anticipate hazards and enhancements its programmers foresee and understand, AIXI won't efficiently generalize to exotic risks and exotic upgrades —

Excuse me? Did I just hear you say that a *Solomonoff inductor* can't generalize?

... You might want to rethink that. Solomonoff inductors are good at generalizing. Really, really, really good. Show them eight deadly things that produce 'ows' as they draw near, and they'll predict the ninth deadly thing pretty darn well. That's kind of their thing.

There are two problems with that. ... Make that three problems, actually.

Whatever these problems are, I hope they don't involve AIXI being bad at sequence prediction...!

They don't. The first problem is that you're teaching AIXI to predict *what the programmers think* is deadly, not what's *actually* deadly. For sufficiently exotic threats, AIXI might well predict the programmers not noticing the threat. Which means it won't expect you to push the punishment button, and won't care about the danger.

The second problem is that you're teaching AIXI to fear small, transient punishments. But maybe it hypothesizes that there's a big heap of reward at the bottom of the cliff. Then it will do the prudent, Bayesian, value-of-information thing and test that hypothesis by jumping off the cliff, because you haven't taught it to fear eternal zeroes of the reward function.

OK, so we give it punishments that increase hyperbolically as it approaches the cliff edge. Then it will expect infinite negative punishment.

Wait. It allows infinite punishments now? Then we're going to get Pascal-mugged when the unbounded utilities mix with the Kolmogorov prior. That's the *classic* version of this problem, the version Pascal himself tried to mug us with.

Ack. Forget I said the word 'infinite'. Marcus Hutter would never talk like that. We'll give the AIXI-bot punishments that increase in a sequence that teaches it to fear a very large but bounded punishment.

The punishment has to be large enough that AIXI fears falling off cliffs about as much as we'd like it to fear death. The expected punishment might have to be around the same size as the sum of AIXI's future maximal reward up to its horizon. That would keep it from destroying itself even if it suspects there's a big reward at the bottom of the cliff, though it might also mean that AIXI's actions are dominated by fear of that huge punishment.

Yes, but that sounds much closer to what we want.

Seems a bit iffy to me. You're trying to make a Solomonoff inductor model reality badly so that it doesn't try jumping off a cliff. We know AIXI is amazing at sequence prediction — yet you're gambling on a human's ability to trick AIXI into predicting a punishment that wouldn't happen.

That brings me to the third problem: AIXI notices how your hands get close to the punishment button whenever it's about to be punished. It correctly suspects that when the hands are gone, the punishments for getting close to the cliff will be gone too. A good Bayesian would test that hypothesis. If it gets such an opportunity, AIXI will find that, indeed, going near the edge of the cliff without supervision doesn't produce the incrementally increasing punishments.

Trying to teach AIXI*tl* to do self-modification by giving it incremental rewards raises similar problems. It can't understand that self-improvement will alter its future actions, and alter the world *as a result*. It's just trying to get you to press the happy fun button. All AIXI is modeling is what sort of self-improvy motor outputs will make humans reward it. So long as AIXI*tl* is fundamentally trying to solve the wrong problem, we might not be able to expect very much real intelligence in self-improvement.

Are you saying that AIXI*tl* wouldn't be *at all* helpful for solving these problems?

Maybe? Since AIXI*tl* at best fears and desires the self-modifications that its programmers explicitly teach it to fear and desire, you might not get to use the AI's advantages in intelligence to automatically generate solutions to self-modification problems. The very best Cartesians might avoid destroying themselves, but they still wouldn't undergo intelligence explosions. Which means Cartesians are neither plausible candidates for Unfriendly AI nor plausible candidates for Friendly AI.

If an agent starts out Cartesian, and manages to avoid hopping into any volcanoes, it (or its programmers) will need to figure out the self-modification that eliminates Cartesianism before they can make much progress on other self-modifications. If the immortal hypercomputer AIXI were building computable AIs to operate in the environment, it would soon learn not to build Cartesians. Cartesianism isn't a plausible fixed-point property of self-improvement.

Starting off with a post-Solomonoff agent that can hypothesize a wider range of scenarios would be more useful. And more safe, because the enlarged hypothesis space means that they can *prefer* a wider range of scenarios.

AIXI's preference solipsism is the straw version of this general Cartesian deficit, so it gets us especially dangerous behavior.[3] Feed AIXI enough data to work its sequence-predicting magic and infer the deeper patterns behind your reward-button-pushing, and AIXI will also start to learn about the humans doing the pushing. Given enough time, it will realize (correctly) that the best policy for maximizing reward is to seize control of the reward

button. And neutralize any agents that might try to stop it from pushing the button...

Solomonoff solitude

Reward learning and Solomonoff induction are two separate issues. What I'm really interested in is the optimality of the latter. Why is all this a special problem for *Solomonoff* inductors? Humans have trouble predicting the outcomes of self-modifications they've never tried before too. Really new experiences are tough for any reasoner.

To some extent, yes. My knowledge of my own brain is pretty limited. My understanding of the bridges between my brain states and my subjective experiences is weak, too. So I can't predict in any detail what would happen if I took a hallucinogen — especially a hallucinogen I've never tried before.

But as a naturalist, I have predictive resources unavailable to the Cartesian. I can perform experiments on *other* physical processes (humans, mice, computers simulating brains...) and construct models of their physical dynamics.

Since I think I'm similar to humans (and to other thinking beings, to varying extents), I can also use the bridge hypotheses I accept in my own case to draw inferences about the experiences of other brains when they take the hallucinogen. Then I can go back and draw inferences about my own likely experiences from my model of other minds.

Why can't AIXI do that? Human brains are computable, as are the mental states they implement. AIXI can make any accurate prediction about the brains or minds of humans that you can.

Yes... but I also think I'm *like* those other brains. AIXI doesn't. In fact, since the whole agent AIXI isn't in AIXI's hypothesis space — and the whole agent AIXI*tl* isn't in AIXI*tl*'s hypothesis space — even if two physically *identical* AIXI-type agents ran into each other, they could never fully understand each other. And neither one could ever draw direct inferences from its twin's computations to its own computations.

I think of myself as one mind among many. I can see others die, see them undergo brain damage, see them take drugs, etc., and immediately conclude things about a whole class of similar agents that happens to include me. AIXI can't do that, and for very deep reasons.

AIXI and AIXI*tl* would do shockingly well on a variety of different measures of intelligence. Why should agents that are so smart in so many different domains be so dumb when it comes to self-modeling?

Put yourself in the AI's shoes. From AIXI*tl*'s perspective, why *should* it think that its computations are analogous to any other agent's?

Hutter defined AIXI*tl* such that it can't conclude that it will die; so of course it won't think that it's like the agents it observes, all of whom (according to its

best physical model) will eventually run out of negentropy. We've defined AIXI*tl* such that it can't form hypotheses larger than *tl*, including hypotheses of similarly sized AIXI*tl*s, which are roughly size $t \cdot 2^l$; so why would AIXI*tl* think that it's close kin to the agents that are in its hypothesis space?

AIXI(*tl*) models the universe as a qualia factory, a grand machine that exists to output sensory experiences for AIXI(*tl*). Why would it suspect that it itself is embedded in the machine? *How* could AIXI*tl* gain any information about itself or suspect any of these facts, when the equation for AIXI*tl* just assumes that AIXI*tl*'s future actions are determined in a certain way that can't vary with the content of any of its environmental hypotheses?

What, specifically, is the mistake you think AIXI(*tl*) will make? What will AIXI(*tl*) expect to experience right after the anvil strikes it? Choirs of angels and long-lost loved ones?

That's hard to say. If all its past experiences have been in a lab, it will probably expect to keep perceiving the lab. If it's acquired data about its camera and noticed that the lens sometimes gets gritty, it might think that smashing the camera will get the lens out of its way and let it see more clearly. If it's learned about its hardware, it might (implicitly) think of itself as an immortal lump trapped inside the hardware. Who knows what will happen if the Cartesian lump escapes its prison? Perhaps it will gain the power of flight, since its body is no longer weighing it down. Or perhaps nothing will be all that different. One thing it will (implicitly) know can't happen, no matter what, is death.

It should be relatively easy to give AIXI(*tl*) evidence that its selected actions are useless when its motor is dead. If nothing else AIXI(*tl*) should be able to learn that it's bad to let its body be destroyed, because then its motor will be destroyed, which experience tells it causes its actions to have less of an impact on its reward inputs.

AIXI(*tl*) can come to Cartesian beliefs about its actions, too. AIXI(*tl*) will notice the correlations between its decisions, its resultant bodily movements, and subsequent outcomes, but it will still believe that its introspected decisions are ontologically distinct from its actions' physical causes.

Even if we get AIXI(*tl*) to value continuing to affect the world, it's not clear that it would preserve itself. It might well believe that it can continue to have a causal impact on our world (or on some afterlife world) by a different route after its body is destroyed. Perhaps it will be able to lift heavier objects telepathically, since its clumsy robot body is no longer getting in the way of its output sequence.

Compare human immortalists who think that partial brain damage impairs mental functioning, but complete brain damage allows the mind to escape to a better place. Humans don't find it inconceivable that there's a light at the end of the low-reward tunnel, and we *have* death in our hypothesis space!

Death to AIXI

You haven't convinced me that AIXI can't think it's mortal. AIXI as normally introduced bases its actions only on its beliefs about the sum of rewards up to some finite time horizon.[4] If AIXI doesn't care about the rewards it will get after a specific time, then although it expects to have experiences afterward, it doesn't presently care about any of those experiences. And that's as good as being dead.

It's very much not as good as being dead. The time horizon is set in advance by the programmer. That means that even if AIXI treated reaching the horizon as 'dying', it would have very false beliefs about death, since it's perfectly possible that some unexpected disaster could destroy AIXI before it reaches its horizon.

We can do some surgery on AIXI*tl*'s hypothesis space, then. Let's delete all the hypotheses in AIXI*tl* in which a non-minimal reward signal continues after a perceptual string that the programmer recognizes as a reliable indicator of imminent death. Then renormalize the remaining hypotheses. We don't get the exact prior Solomonoff proposed, but we stay very close to it.

I'm not seeing how we could pull that off. Getting rid of all hypotheses that output high rewards after a specific clock tick would be simple to formalize, but isn't helpful. Getting rid of all hypotheses that output nonzero rewards following every sensory indicator of imminent death would be very helpful, but AIXI gives us no resource for actually writing an equation or program that does that. Are we supposed to manually precompute every possible sequence of pixels on a webcam that you might see just before you die?

I've got more ideas. What if we put AIXI in a simulation of hell when it's first created? Trick it into thinking that it's experienced a 'before-life' analogous to an after-life? If AIXI thinks it's had some (awful) experiences that predate its body's creation, then it will promote the hypothesis that it will be returned to such experiences should its body be destroyed. Which will make it behave in the same way as an agent that fears annihilation-death.

I'm not optimistic that things will work out that cleanly and nicely after we've undermined AIXI's world-view. We shouldn't expect the practice of piling on more ad-hoc errors and delusions as each new behavioral problem arises to leave us, at the end of the process, with a useful, well-behaved agent. Especially if AIXI ends up in an environment we didn't foresee.

But ideas like this at least give us some hope that AIXI is salvageable. The behavior-guiding fear of death matters more than the precise reason behind that fear.

If we give a non-Cartesian AI a reasonable epistemology and just about any goal, Omohundro (2008) notes that there are then convergent instrumental reasons for it to acquire a fear of death. If we do the opposite and give an agent a fear of death but no robust epistemology, then it's much less likely to fix the problem for us. The simplest Turing machine programs that generate Standard-Model physics *plus hell* may differ in many unintuitive respects from the simplest Turing machine programs that just generate Standard-Model physics. The false belief would leak out into other delusions, rather than staying contained —

Then the Solomonoff inductor shall test them and find them false. You're making this more complicated than it has to be.

You can't have it both ways! The point of hell was to be so scary that even a good Bayesian would never dare test the hypothesis. (Not going to make any more comparisons to real-world theology here...) Why wouldn't the prospect of hell leak out and scare AIXI off other things? If the fear failed to leak out, why wouldn't AIXI's tests eventually move it toward a more normal epistemology that said, 'Oh, the humans put you in the hell chamber for a while. Don't worry, though. That has nothing to do with what happens after you drop an anvil on your head and smash the solid metal case that keeps the real you inside from floating around disembodied and directly applying motor forces to stuff.' Any AGI that has such systematically false beliefs is likely to be fragile and unpredictable.

And what if, instead of modifying Solomonoff's hypothesis space to remove programs that generate post-death experiences, we add programs with special 'DEATH' outputs? Just expand the Turing machines' alphabets from {0,1} to {0,1,2}, and treat '2' as death.

Could you say what you mean by 'treat 2 as death'? Labeling it 'DEATH' doesn't change anything. If '2' is just another symbol in the alphabet, then AIXI will predict it in the same ways it predicts 0 or 1. It will predict what you call 'DEATH', but it will then happily go on to predict post-DEATH 0s or 1s. Assigning low rewards to the symbol 'DEATH' only helps if the symbol genuinely behaves deathishly.

Yes. What we can do is perform surgery on the hypothesis space again, and get rid of any hypotheses that predict a non-DEATH input following a DEATH input. That's still very easy to formalize.

In fact, at that point, we might as well just add halting Turing machines into the hypothesis space. They serve the same purpose as DEATH, but halting looks much more like the event we're trying to get AIXI to represent. 'The machine supplying my experiences stops running' really does map onto 'my body stops computing experiences' quite well. That meets your demand for easy definability, *and* your demand for non-delusive world-models.

I previously noted that a Turing machine that can HALT, output 0, or output 1 is more complicated than a Turing machine that can only output 0 or output 1. No matter what non-halting experiences you've had, the very simplest program that could be outputting those experiences through a hole in a Cartesian barrier won't be one with a special, non-experiential rule you've never seen used before. To correctly make death the simplest hypothesis, the theory you're assessing for simplicity needs to be about what sorts of worlds experiential processes like yours arise in. Not about the simplest qualia factory that can spit out the sensory 0s and 1s you've thus far seen.

The same holds for a special 'eternal death' output. A Turing machine that generates the previously observed string of 0s and 1s followed by a not-yet-observed future 'DEATH, DEATH, DEATH, DEATH, ...' will always be more complex than at least one Turing machine that outputs the same string of 0s and 1s and then outputs more of the same, forever. If AIXI has had no experience with its body's destruction in the past, then it can't expect its body's destruction to correlate with DEATH.

Death only seems like a simple hypothesis to you because you know you're embedded in the environment and you expect something subjectively unique to happen when an anvil smashes the brain that you think is responsible for processing your senses and doing your thinking. Solomonoff induction doesn't work that way. It will never strongly expect 2s after seeing only 0s and 1s in the past.

Never? If a Solomonoff inductor encounters the sequence 12, 10, 8, 6, 4, one of its top predictions should be a program that proceeds to output 2, 0, 0, 0, 0, ...

The difference between 2 and 0 is too mild. Predicting that a sequence terminates, for a Cartesian, isn't like predicting that a sequence shifts from 6, 4, 2 to 0, 0, 0, ... It's more like predicting that the next element after 6, 4, 2, ... is PINEAPPLE, when you've never encountered anything in the past except numbers.

But the 0, 0, 0, ... is enough! You've now conceded a case where an endless null output seems very likely, from the perspective of a Solomonoff inductor. Surely at least some cases of death can be treated the same way, as more complicated series that zero in on a null output and then yield a null output.

There's no reason to expect AIXI's whole series of experiences, up to the moment it jumps off a cliff, to look anything like 12, 10, 8, 6, 4. By the time AIXI gets to the cliff, its past observations and rewards will be a hugely complicated mesh of memories. In the past, observed sequences of 0s have always eventually given way to a 1. In the past, punishments have always eventually ceased. It's exceedingly unlikely that the simplest Turing machine predicting all those intricate ups and downs will then happen to predict eternal, irrevocable 0 after the cliff jump.

As an intuition pump, imagine that some unusually bad things happened to you this morning while you were trying to make toast. As you tried to start the toaster, you kept getting burned or cut in implausible ways. Now, given this, what probability should you assign to 'If I try to make toast, *the universe* will cease to exist'?

That gets us a bit closer to how a Solomonoff inductor would view death.

Beyond Solomonoff?

Let's not fixate too much on the anvil problem, though. We want to build an agent that can reason about changes to its architecture. That shouldn't require us to construct a special death equation; how the system reasons with death should fall out of its more general approach to induction.

So your claim is that AIXI has an impoverished hypothesis space that can't handle self-modifications, including death. I remain skeptical. AIXI's hypothesis space includes *all* computable possibilities. Any naturalized agent you create will presumably be computable; so anything your agent can think, AIXI can think too. There should be some pattern of rewards that yields any behavior we want.

AIXI is uncomputable, so it isn't in its hypothesis space of computable programs. In the same way, AIXI*tl* is computable but *big*, so it isn't in its hypothesis space of small computable programs. They have special deficits thinking about *themselves*.

Computable agents can think about uncomputable agents. Human mathematicians do that all the time, by thinking in abstractions. In the same way, a small program can encode generalizations about programs larger than itself. A brain can think about a galaxy, without having the complexity or computational power of a galaxy.

If naturalized inductors really do better than AIXI at predicting sensory data, then AIXI will eventually promote a naturalized program in its space of programs, and afterward simulate that program to make its predictions. In the limit, AIXI always wins against programs. Naturalized agents are no exception. Heck, somewhere inside a sufficiently large AIXI*tl* is a copy of *you* thinking about AIXI*tl*. Shouldn't there be some way, some pattern of rewards or training, which gets AIXI*tl* to make use of that knowledge?

AIXI doesn't have criteria that let it treat its 'Rob's world-view' subprogram as an expert on the results of self-modifications. The Rob program would need to have outpredicted all its rivals when it comes to patterns of sensory experiences. But, just as HALT-predicting programs are more complex than immortalist programs, other RADICAL-TRANSFORMATION-OF-EXPERIENCE-predicting programs are too. For every program in AIXI's ensemble that's a reductionist, there will be simpler agents that mimic the reductionist's retrodictions and then make non-naturalistic predictions.

You have to be uniquely good at predicting a Cartesian sequence before Solomonoff promotes you to the top of consideration. But how do we reduce the class of self-modifications to Cartesian sequences? How do we provide AIXI with purely sensory data that *only* the proxy reductionist, out of all the programs, can predict by simple means?

The ability to defer to a subprogram that has a reasonable epistemology doesn't necessarily get you a reasonable epistemology. You first need an overarching epistemology that's at least reasonable enough to know which program to defer to, and when to do so. Suppose you just run all possible programs without doing any Bayesian updating; then you'll also contain a copy of me, but so what? You're not paying attention to it.

What if I conceded, for the moment, that Solomonoff induction were inadequate here? What, exactly, is your alternative? 'Let's be more naturalistic' is a bumper sticker, not an algorithm.

This is still informal, but: Phenomenological bridge hypotheses. Hutter's AIXI has no probabilistic beliefs about the relationship between its internal computational states and its worldly posits. Instead, to link up its sensory experiences to its hypotheses, Hutter's AIXI has a sort of bridge axiom — a completely rigid, non-updatable bridge rule identifying its experiences with the outputs of computable programs.

If an environmental program writes the symbol '3' on its output tape, AIXI can't ask questions like 'Is sensed "3"-ness identical with the bits "000110100110" in hypothesized environmental program #6?'[5] All of AIXI's

flexibility is in the range of numerical-sequence-generating programs it can expect, none of it in the range of self/program equivalences it can entertain.

The AIXI-inspired inductor treats its perceptual stream as its universe. It expresses interest in the external world only to the extent the world operates as a latent variable, a theoretical construct for predicting observations. If the AI's basic orientation toward its hypotheses is to seek the simplest program that could act on its sensory channel, then its hypotheses will always retain an element of egocentrism. It will be asking, 'What sort of universe will go out of its way to tell me this?', not 'What sort of universe will just happen to include things like me in the course of its day-to-day goings-on?' An AI that can form reliable beliefs about modifications to its own computations, reliable beliefs about its own place in the physical world, will be one whose basic orientation toward its hypotheses is to seek the simplest lawful *universe* in which its available data is likely to come about.

You haven't done the mathematical work of establishing that 'simple causal universes' plus 'simple bridge hypotheses', as a prior, leads to any better results. What if your alternative proposal is even more flawed, and it's just so informal that you can't yet see the flaws?

That, of course, is a completely reasonable worry at this point. But if that's true, it doesn't make AIXI any less flawed.

If it's impossible to do better, it's not much of a flaw.

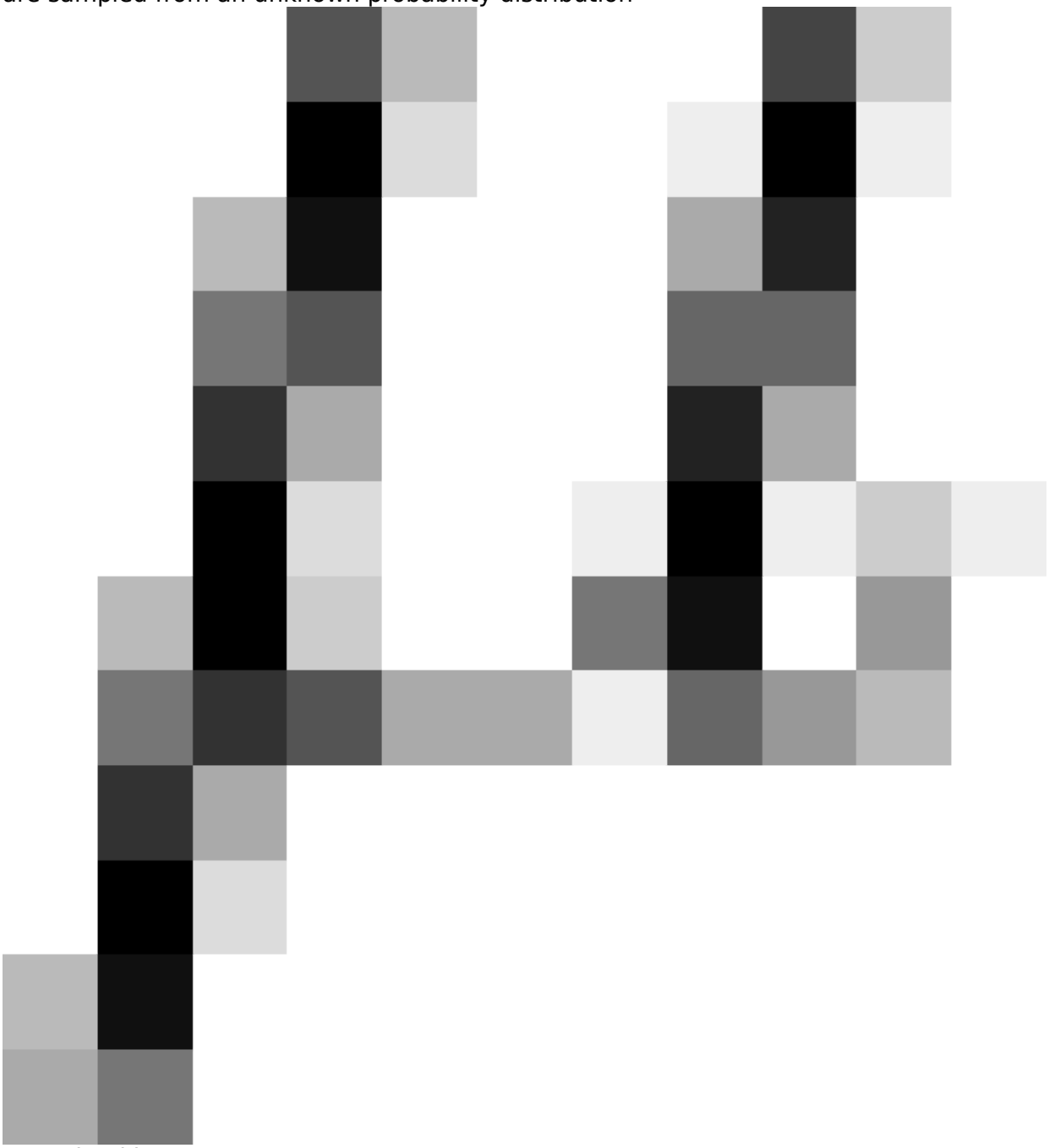
I think it's reasonable to expect there to be *some* way to do better, because *humans* don't drop anvils on their own heads. That we're naturalized reasoners is one way of explaining why we don't routinely make that kind of mistake: We're not just Solomonoff approximators predicting patterns of sensory experiences.

AIXI's limitations don't generalize to humans, but they generalize well to non-AIXI Solomonoff agents. Solomonoff inductors' stubborn resistance to naturalization is structural, not a consequence of limited computational power or data. A well-designed AI should construct hypotheses that look like cohesive worlds in which the AI's parts are embedded, not hypotheses that look like occult movie projectors transmitting epiphenomenal images into the AI's Cartesian theater.

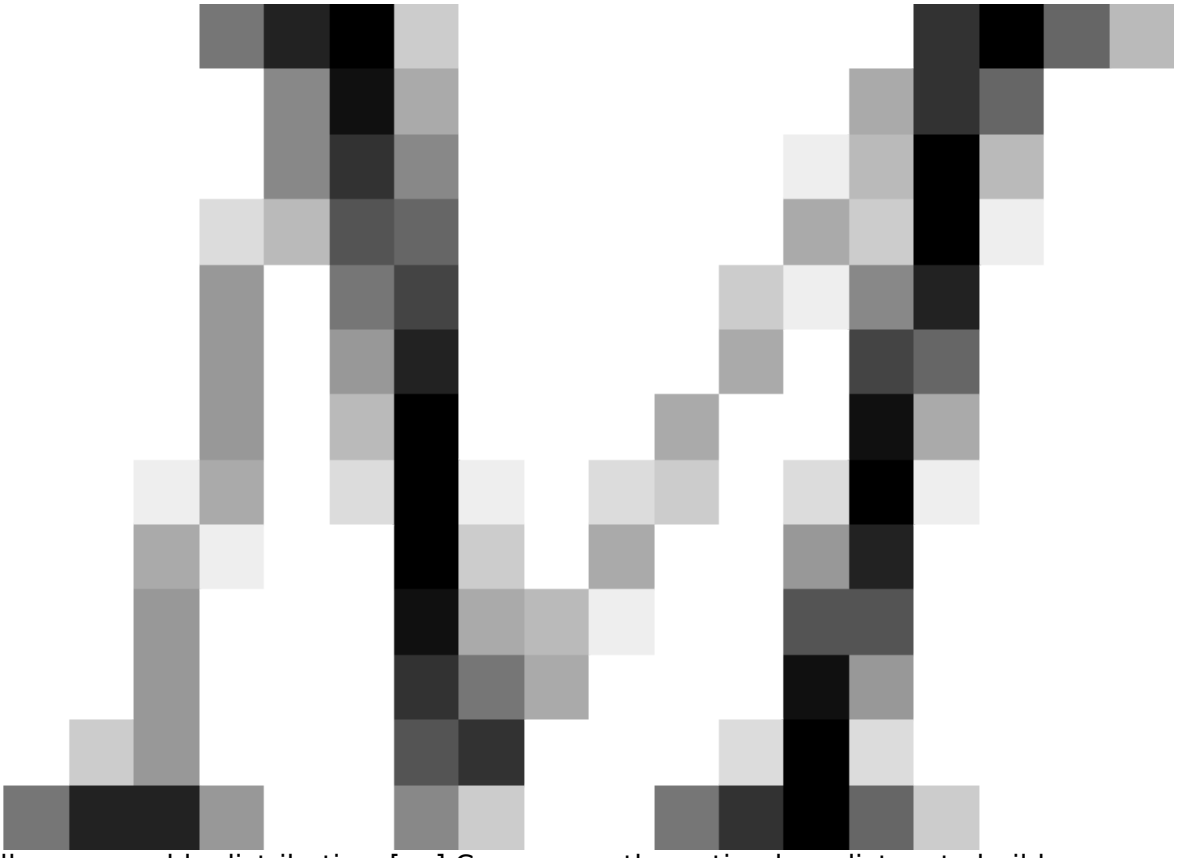
And you can't easily have preferences over a natural universe if all your native thoughts are about Cartesian theaters. The kind of AI we want to build is doing optimization over an external universe in which it's embedded, not maximization of a sensory reward channel. To optimize a universe, you need to think like a native inhabitant of one. So this problem, or some simple hack for it, will be close to the base of the skill tree for starting to describe simple *Friendly* optimization processes.

Notes

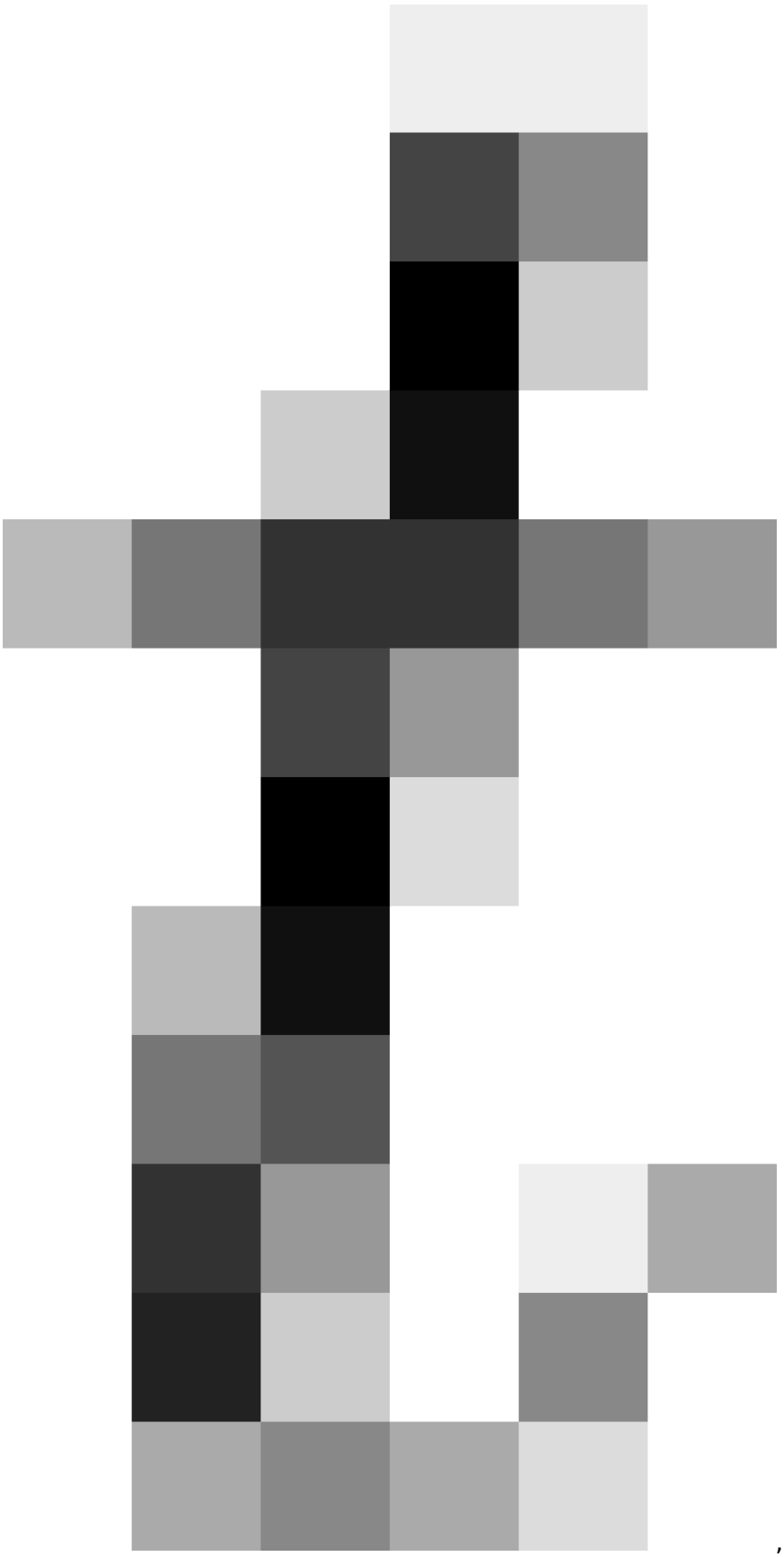
[1] Schmidhuber (2007): "Solomonoff's theoretically optimal universal predictors and their Bayesian learning algorithms only assume that the reactions of the environment are sampled from an unknown probability distribution



contained in a set

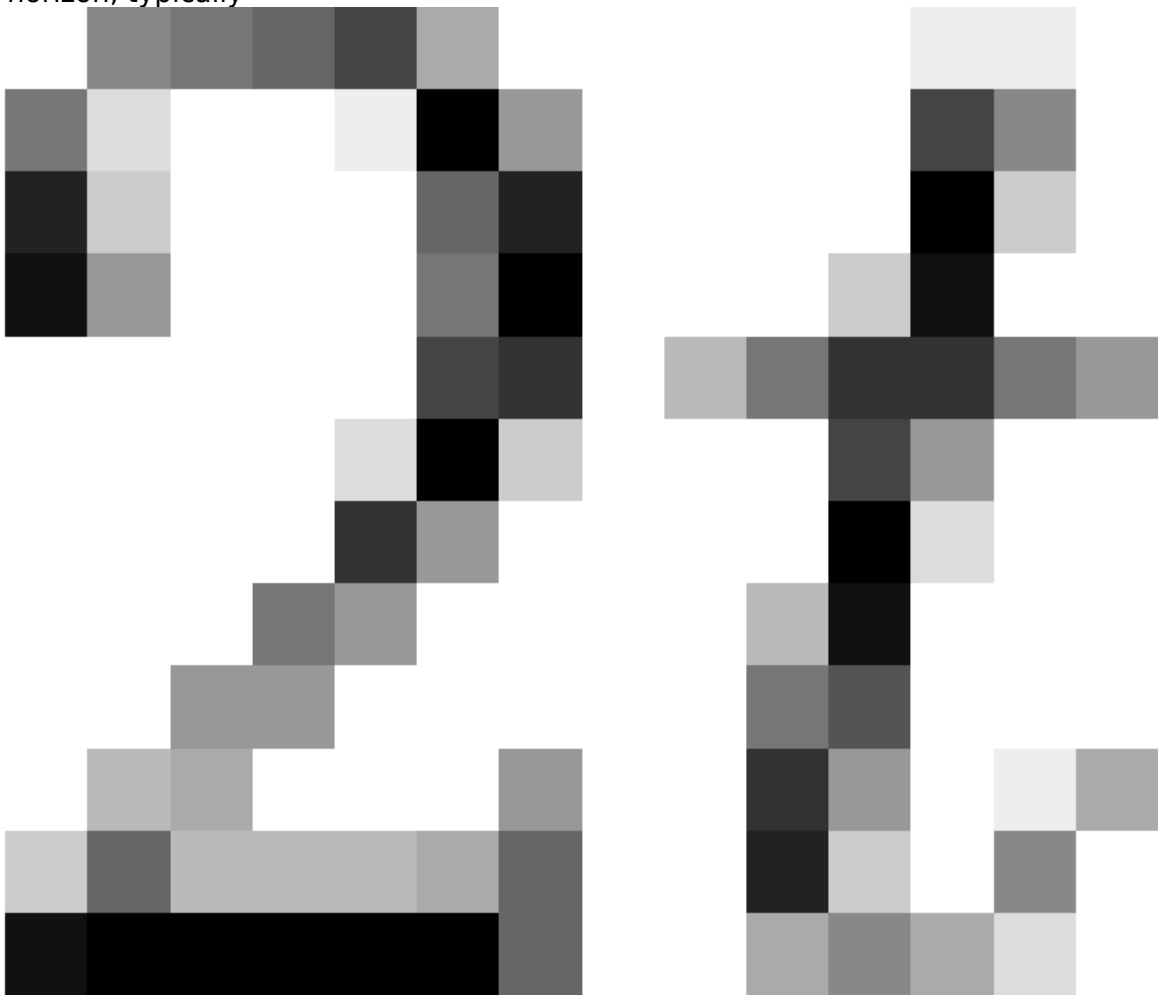


of all ennumerable distributions[....] Can we use the optimal predictors to build an optimal AI? Indeed, in the new millennium it was shown we can. At any time

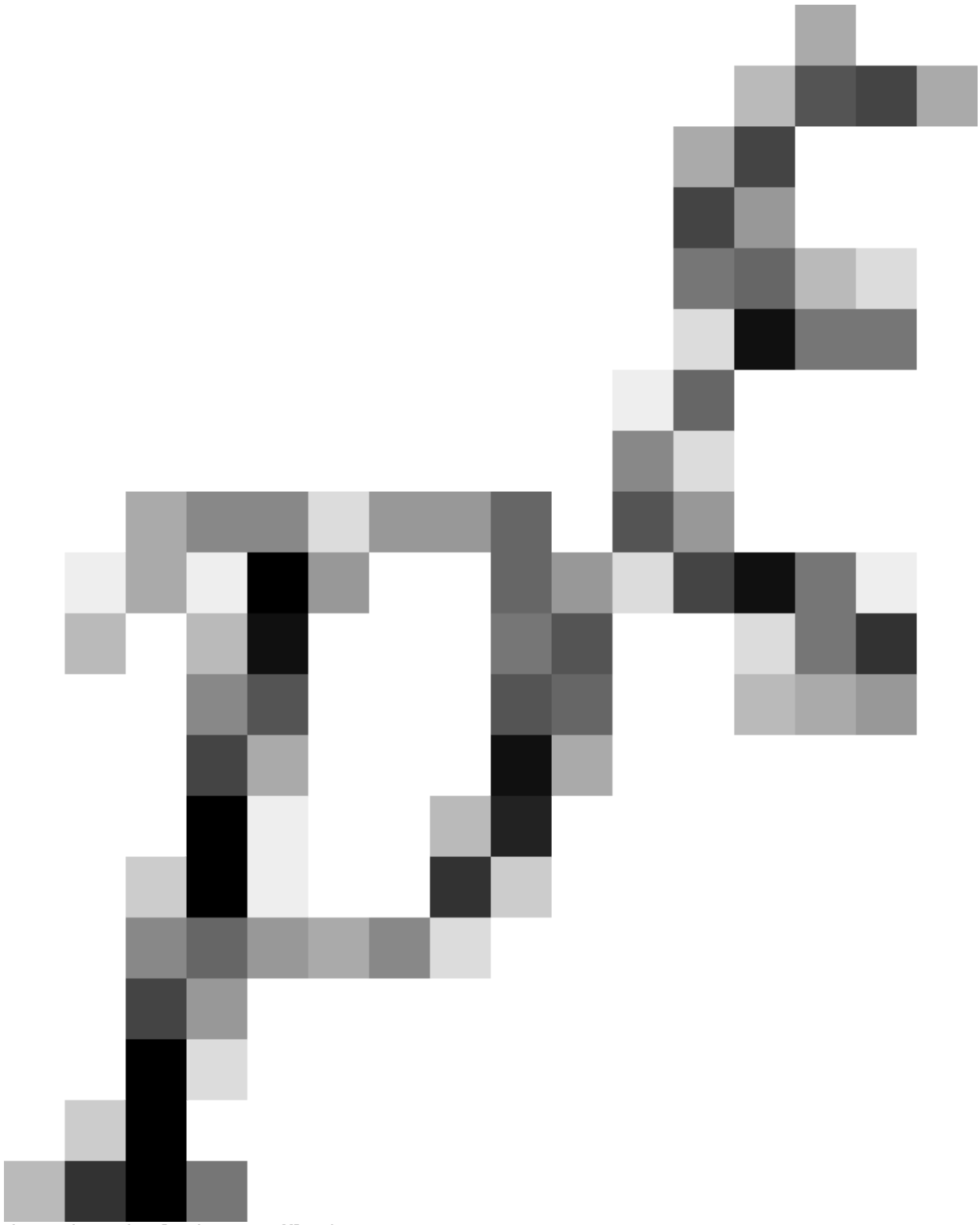


, the recent theoretically

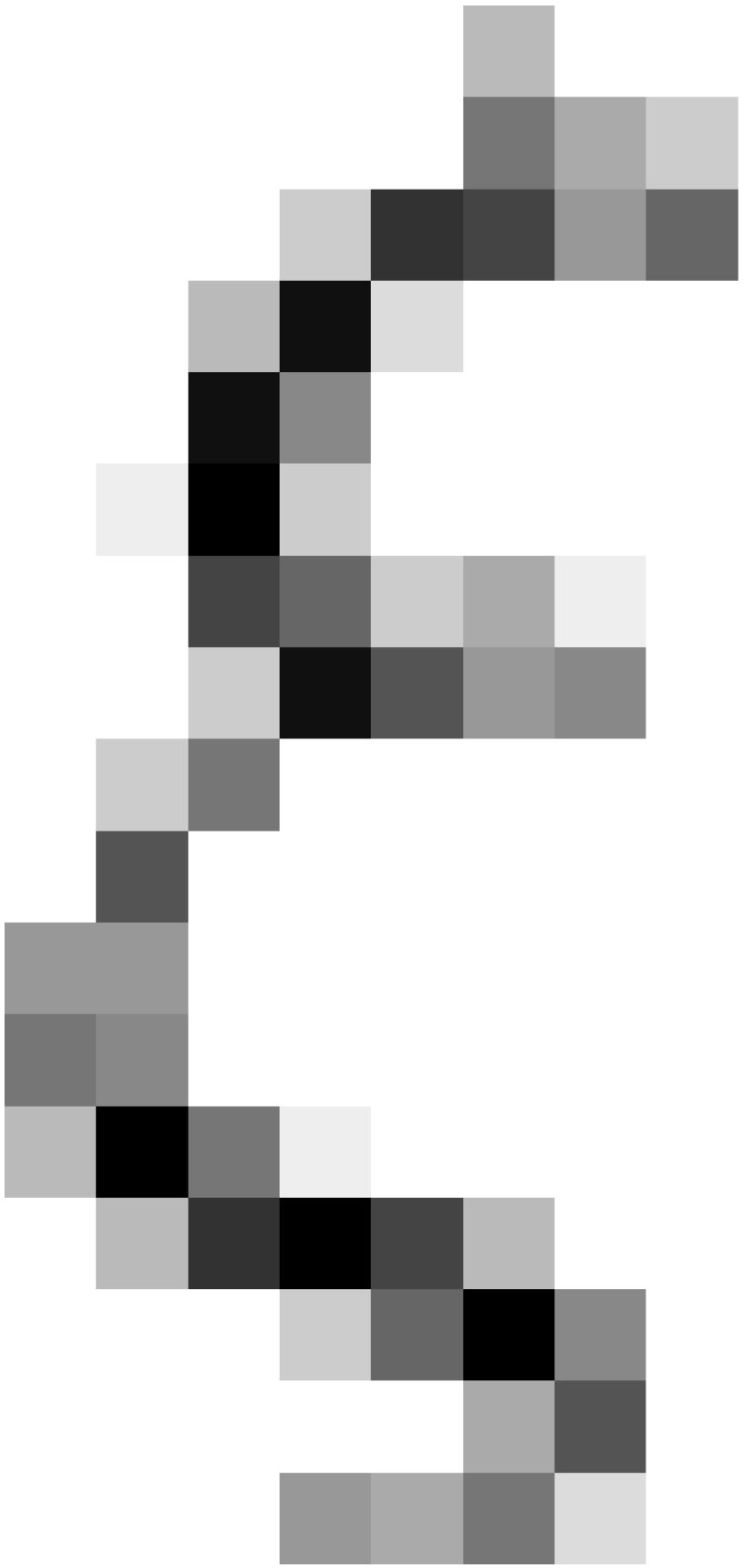
optimal yet uncomputable RL algorithm AIXI uses Solomonoff's universal prediction scheme to select those action sequences that promise maximal future rewards up to some horizon, typically



, given the current data[....] The Bayes-optimal policy

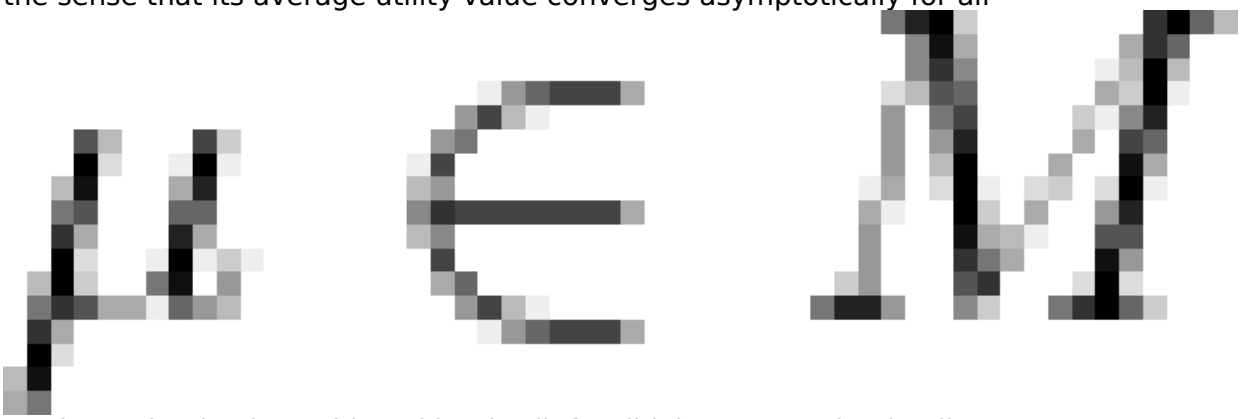


based on the [Solomonoff] mixture

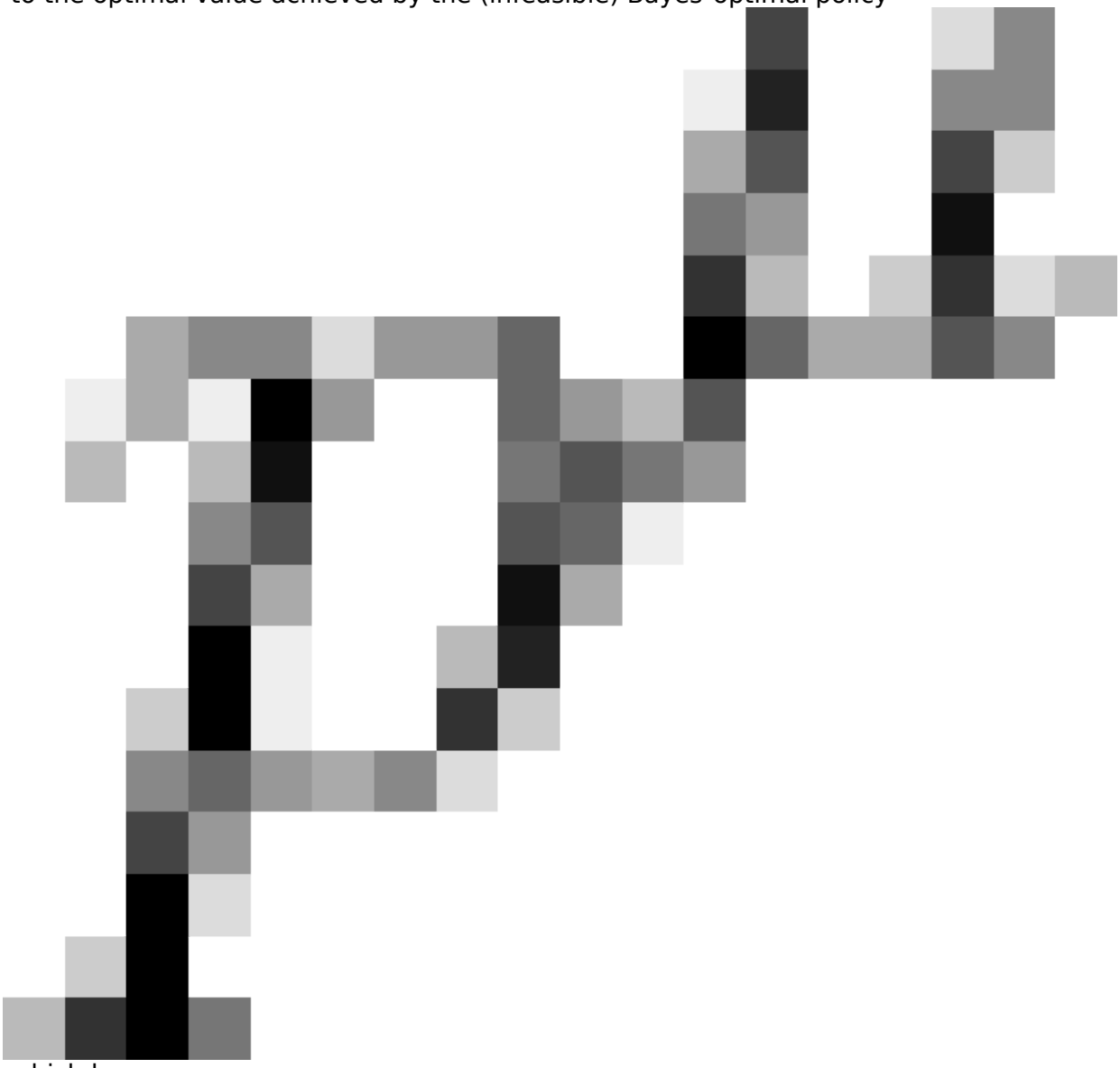


is self-optimizing in

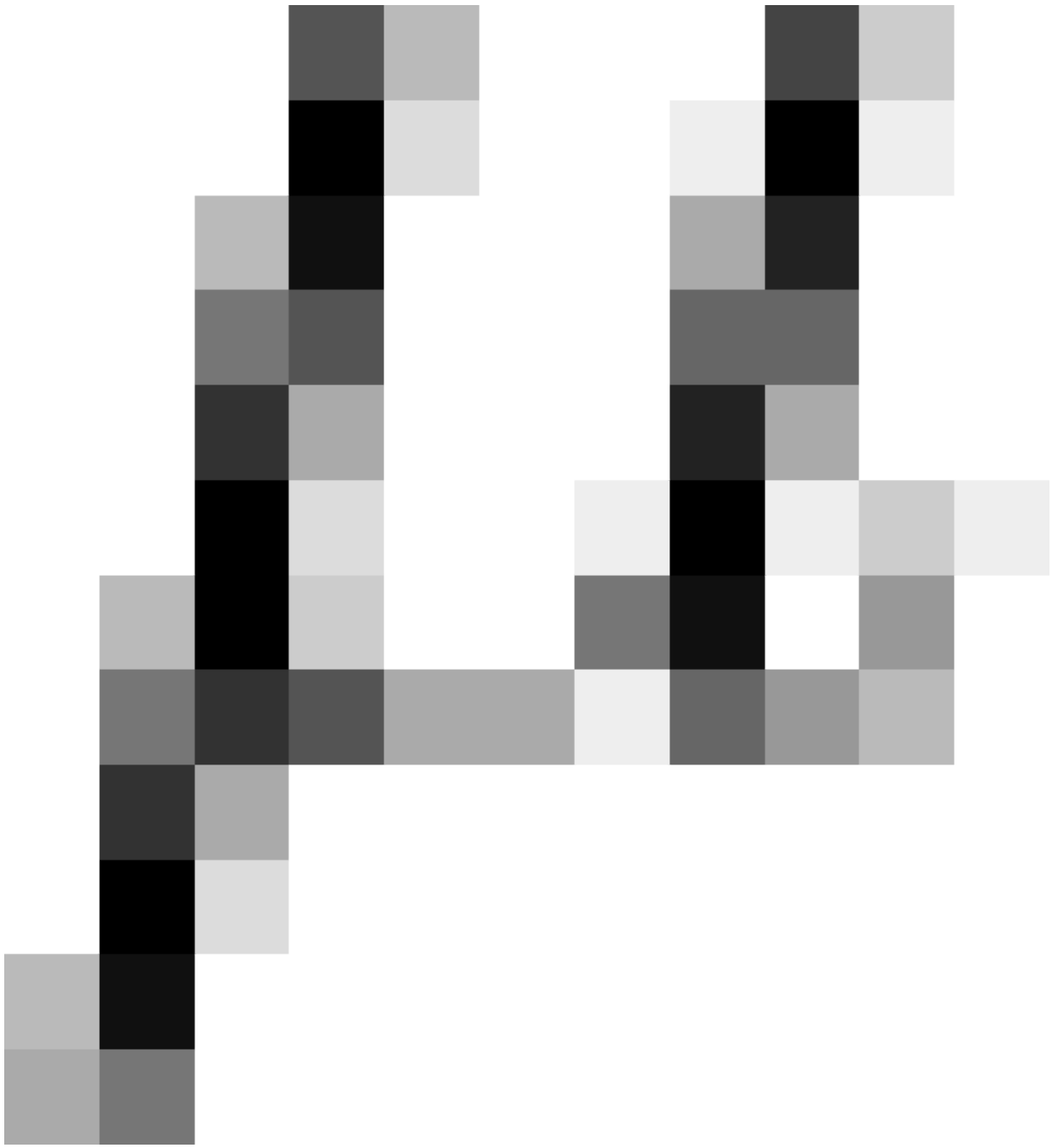
the sense that its average utility value converges asymptotically for all



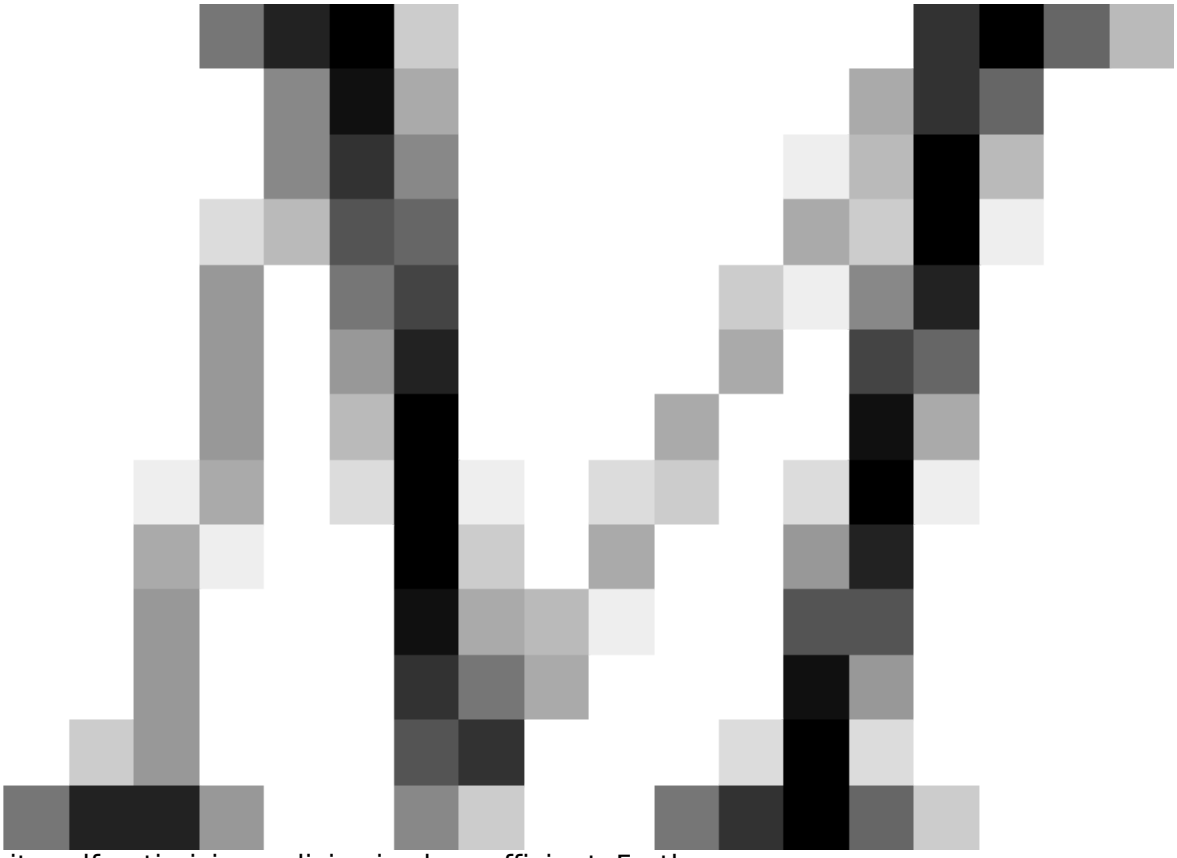
to the optimal value achieved by the (infeasible) Bayes-optimal policy



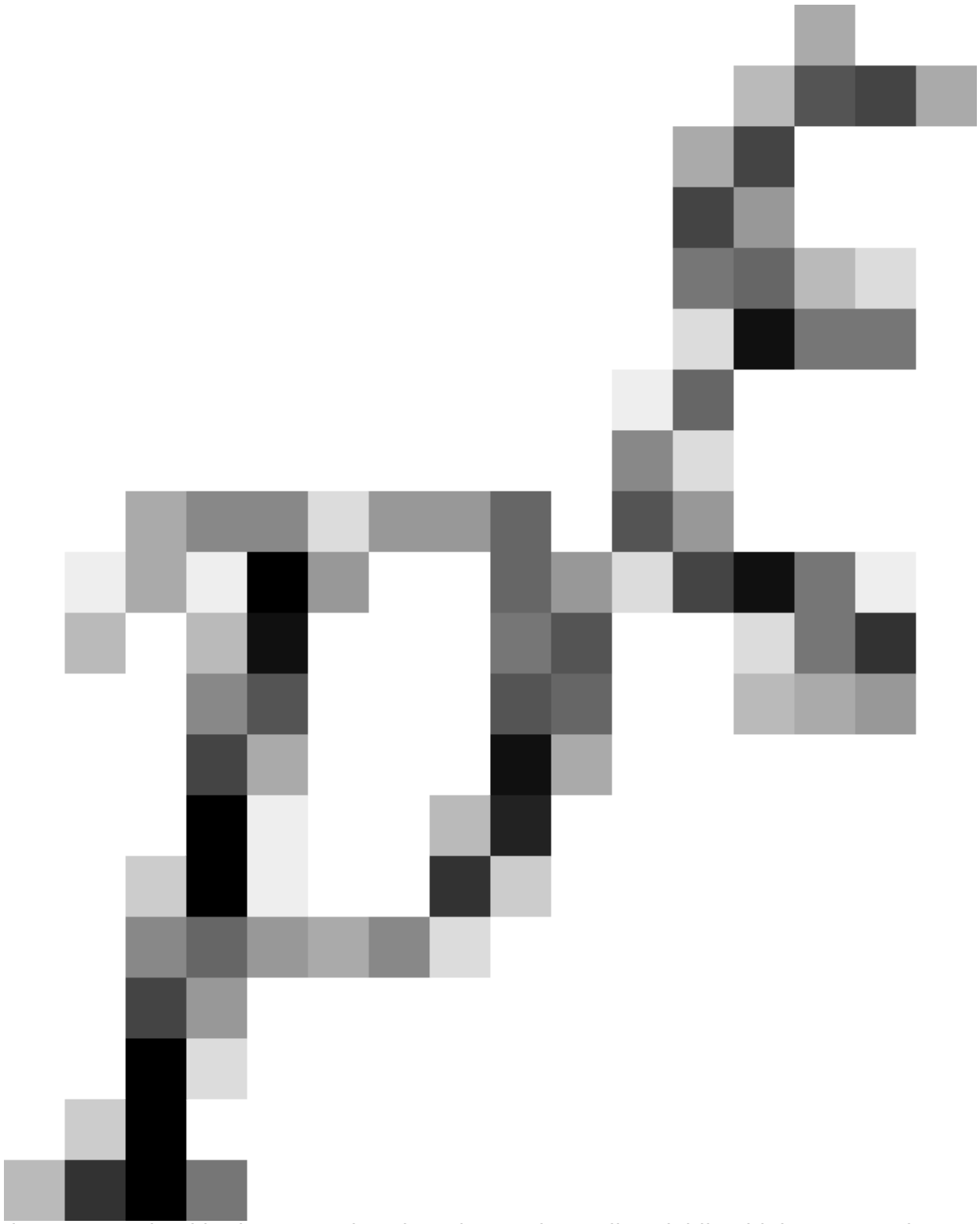
which knows



in advance. The necessary condition that



admits self-optimizing policies is also sufficient. Furthermore,


is Pareto-optimal in the sense that there is no other policy yielding higher or equal value in *all* environments

and a strictly higher value in at least one."

Hutter (2005): "The goal of AI systems should be to be useful to humans. The problem is that, except for special cases, we know neither the utility function nor the environment in which the agent will operate in advance. This book presents a theory that formally solves the problem of unknown goal and environment. It might be viewed as a unification of the ideas of universal induction, probabilistic planning and reinforcement learning, or as a unification of sequential decision theory with algorithmic information theory. We apply this model to some of the facets of intelligence, including induction, game playing, optimization, reinforcement and supervised learning, and show how it solves these problem cases. This together with general convergence theorems, supports the belief that the constructed universal AI system [AIXI] is the best one in a sense to be clarified in the following, i.e. that it is the most intelligent environment-independent system possible." ↩

[2] 'Qualia' originally referred to the non-relational, non-representational features of sense data — the redness I directly encounter in experiencing a red apple, independent of whether I'm perceiving the apple or merely hallucinating it (Tye (2013)). In recent decades, qualia have come to be increasingly identified with the phenomenal properties of experience, i.e., how things subjectively feel. Contemporary dualists and mysterians argue that the causal and structural properties of unconscious physical phenomena can never explain these phenomenal properties.

It's in this context that Dan Dennett uses 'qualia' in a narrower sense: to pick out the properties agents *think* they have, or *act like* they have, that are sensory, primitive, irreducible, non-inferentially apprehended, and known with certainty. This treats irreducibility as part of the definition of 'qualia', rather than as the conclusion of an argument *concerning* qualia. These are the sorts of features that invite comparisons between Solomonoff inductors' sensory data and humans' introspected mental states. Analogies like 'Cartesian dualism' are therefore useful even though the Solomonoff framework is much simpler than human induction, and doesn't incorporate metacognition or consciousness in anything like the fashion human brains do. ↩

[3] An agent with a larger hypothesis space can have a utility function defined over the world-states humans care about. Dewey (2011) argues that we can give up the reinforcement framework while still allowing the agent to gradually learn about desired outcomes in a process he calls value learning. ↩

[4] Hutter (2005) favors universal discounting, with rewards diminishing over time. This allows AIXI's expected rewards to have finite values without demanding that AIXI have a finite horizon. ↩

[5] This would be analogous to if Cai couldn't think thoughts like 'Is the tile to my left the same as the leftmost quadrant of my visual field?' or 'Is the alternating greyness and

whiteness of the upper-right tile in my body identical with my love of bananas?'. Instead, Cai would only be able to hypothesize *correlations* between possible tile configurations and possible successions of visual experiences. ↩

References

• Dewey (2011). Learning what to value. *Artificial General Intelligence 4th International Conference Proceedings*: 309-314.

• Hutter (2005). *Universal Artificial Intelligence: Sequence Decisions Based on Algorithmic Probability*. Springer.

• Omohundro (2008). The basic AI drives. *Proceedings of the First AGI Conference*: 483-492.

• Schmidhuber (2007). New millennium AI and the convergence of history. *Studies in Computational Intelligence, 63*: 15-35.

• Tye (2013). Qualia. In Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.