Tiling Agents for Self-Modifying AI, and the Löbian Obstacle*

Yudkowsky, Eliezer Herreshoff, Marcello

October 7, 2013

(Early Draft)

Abstract

We model self-modification in AI by introducing "tiling" agents whose decision systems will approve the construction of highly similar agents, creating a repeating pattern (including similarity of the offspring's goals). Constructing a formalism in the most straightforward way produces a Gödelian difficulty, the "Löbian obstacle." By technical methods we demonstrate the possibility of avoiding this obstacle, but the underlying puzzles of rational coherence are thus only partially addressed. We extend the formalism to partially unknown deterministic environments, and show a very crude extension to probabilistic environments and expected utility; but the problem of finding a fundamental decision criterion for self-modifying probabilistic agents remains open.

1 Introduction

Suppose a sufficiently intelligent agent possesses a goal (or a preference ordering over outcomes, or a utility function over probabilistic outcomes). One possible means for the agent to achieve its goal is to construct another agent that shares the same goal (Omohundro 2008; Bostrom 2012).[1] As a special case of agents constructing successors with equivalent goals, a machine intelligence

*The research summarized in this paper was conducted first by Yudkowsky and Herreshoff; refined at the November 2012 MIRI Workshop on Logic, Probability, and Reflection with Paul Christiano and Mihaly Barasz; and further refined at the April 2013 MIRI Workshop on Logic, Probability, and Reflection with Paul Christiano, Benja Fallenstein, Mihaly Barasz, Patrick LaVictoire, Daniel Dewey, Qiaochu Yuan, Stuart Armstrong, Jacob Steinhardt, Jacob Taylor, and Andrew Critch.

1. If you wanted a road to exist to a certain city, and your capabilities included the construction of other creatures, you might choose to construct a creature which also wanted a road to exist to that city and would accordingly seek to build one, so that its efforts would tend to accomplish your original goal.

may wish to change its own source code—self-improve. This can be viewed as constructing a successor agent that shares most of your program code and runs on the same hardware as you, then replacing yourself with that agent.[2] We shall approach the subject of AI self-modification and self-improvement by considering it as a special case of agents constructing other agents with similar preferences.

In a self-modifying AI, most self-modifications should not change most aspects of the AI; it would be odd to consider agents that could only make large, drastic self-modifications. To reflect this desideratum within the viewpoint from agents constructing other agents, we will examine agents which construct successor agents of highly similar design, so that the sequence of agents "tiles" like a repeating pattern of similar shapes on a tiled floor.

In attempting to describe agents whose decision criteria approve the construction of highly similar agents, we shall encounter a Gödelian difficulty[3] in the form of Löb's Theorem: A consistent mathematical theory $\mathcal{T}$ cannot trust itself in the sense of verifying that a proof in $\mathcal{T}$ of any formula ϕ implies ϕ's truth—we cannot have the self-verifying scheme (over all formulas ϕ) of $\forall\phi : \mathcal{T} \vdash \Box_{\mathcal{T}}\ulcorner\phi\urcorner \rightarrow \phi$ (Löb 1955).[4] We shall construct a natural-seeming schema for agents reasoning about other agents, which will at first seem to imply that such an agent can only trust the reasoning of successors that use weaker mathematical systems than its own. This in turn would imply that an agent architecture can only tile a finite chain of successors (or make a finite number of self-modifications) before running out of "trust." This Gödelian difficulty poses challenges both to the construction of successors, and to a reflective

2. If you wanted a road to a certain city to exist, you might try attaching more powerful arms to yourself so that you could lift paving stones into place. This can be viewed as a special case of constructing a new creature with similar goals and more powerful arms, and then replacing yourself with that creature.

3. That human beings are computable does not imply that Gödelian-type difficulties will never present themselves as problems for AI work; rather it implies that any such Gödelian difficulty ought to be equally applicable to a human, and that any human way of bypassing the Gödelian difficulty could presumably be carried over to an AI. E.g., the halting theorem imposes limits on computable AIs, imposes limits on humans, and will presumably impose limits on any future intergalactic civilization; however, the observed existence of humans running on normal physics implies that human-level cognitive intelligence does not require solving the general halting problem. Thus, any supposed algorithm for general intelligence which demands a halting oracle is not revealing the uncomputability of humans, but rather is making an overly strong demand. It is in this spirit that we will investigate, and attempt to deal with, the Gödelian difficulty exposed in one obvious-seeming formalism for self-modifying agents.

4. $\ulcorner\phi\urcorner$ denotes the Gödel number of the formula ϕ, and $\Box_{\mathcal{T}}\ulcorner\phi\urcorner$ denotes the proposition that there exists an object p such that p is the Gödel number of a proof in $\mathcal{T}$ of $\ulcorner\phi\urcorner$. E.g., letting $\mathcal{PA}$ represent the system of first-order Peano Arithmetic, $\Box_{\mathcal{PA}}\ulcorner S0 + S0 = SS0\urcorner$ stands for the proposition $\exists p : Bew_{\mathcal{PA}}(p, \ulcorner S0 + S0 = SS0\urcorner)$ where $Bew_{\mathcal{PA}}(p, \ulcorner\phi\urcorner)$ is a formula (in Δ_0) stating that p Gödel-encodes a proof of the quoted theorem $\ulcorner\phi\urcorner$ from the axioms of Peano Arithmetic. Thus $\mathcal{PA} \vdash \Box_{\mathcal{PA}}\ulcorner S0 + S0 = SS0\urcorner$ states (accurately) that first-order arithmetic proves that there exists a proof in $\mathcal{PA}$ that $1 + 1 = 2$. Also, whenever a quantifier over formulas $\forall\phi$ appears, this denotes a meta-language schema with a separate axiom or theorem in the object language for each formula ϕ.

agent's immediate self-consistency.

We shall present several different techniques to bypass this Gödelian difficulty, demonstrating indefinitely tiling sequences of agents maintaining trust in the same mathematical system. Some unachieved desiderata of reflective coherence will remain; and while the technical methods used will demonstrate the technical possibility, they will not be plausible as basic structures of rational agents. We shall then pass from known environments to partially known environments within the formalism, and make a preliminary attempt to pass from logical agents to probabilistic agents that calculate expected utility. Some flaws and unfortunate-seeming properties of the currently proposed formalism will also be discussed. The ultimate problem of proposing a satisfactory fundamental decision criterion for self-modifying agents remains open.[5]

2 Logical agents that construct successors

Suppose an agent $\mathcal{A}^1$ with a satisficing goal $\mathcal{G}$—any outcome shall either satisfy or not satisfy $\mathcal{G}$, and $\mathcal{A}^1$'s sole preference is for outcomes satsifying $\mathcal{G}$ over outcomes not satisfying $\mathcal{G}$. We shall initially suppose that the agent $\mathcal{A}^1$ occupies a crisp, fully known, deterministic, closed environment; then first-order logic will be a good representational fit for reasoning about this environment. Suppose also that the environment contains "transistors," objects which can be configured by $\mathcal{A}^1$ to perform computations. Then $\mathcal{A}^1$ might choose to use these transistors to build an offspring $\mathcal{A}^0$ which also shares the goal $\mathcal{G}$.

Suppose $\mathcal{A}^1$ constructs $\mathcal{A}^0$ as a satisficing logical agent that only takes actions that $\mathcal{A}^0$ can prove to achieve the goal $\mathcal{G}$ (shared by $\mathcal{A}^1$ and $\mathcal{A}^0$). We shall write $\mathcal{A}^0$'s *criterion of action*—the conditions under which $\mathcal{A}^0$ takes an action—using the notation:

$$\bar{b}_i \quad \Rightarrow \quad \mathcal{A}^0 \Vdash \bar{b}_i \rightarrow \mathcal{G}, \tag{2.1}$$

where $\bar{b}_i$ is a proposition denoting the actual performance of an action b_i within $\mathcal{A}^0$'s range of action ($b_i \in Acts^0$), and $\Vdash$ is a special turnstile indicating that an agent has cognitively concluded a belief. Hence (2.1) states that the actual performance of any action b_i only takes place when $\mathcal{A}^0$ has cognitively proven a theorem that the performance of b_i implies $\mathcal{G}$. Without sensors (section 6), $\mathcal{A}^0$ can conclude a belief ϕ only if ϕ is a consequence of its axioms Axm^0:

$$\mathcal{A}^0 \Vdash \phi \quad \Rightarrow \quad Axm^0 \vdash \phi \quad \text{(without sensors)}. \tag{2.2}$$

5. The paper is meant to be interpreted within an agenda of "Begin tackling the conceptual challenge of describing a stably self-reproducing decision criterion by inventing a simple formalism and confronting a crisp difficulty"; not as "We think this Gödelian difficulty will block AI", nor "This formalism would be good for an actual AI", nor "A bounded probabilistic self-modifying agent would be like this, only scaled up and with some probabilistic and bounded parts tacked on". Similarly we use first-order logic because reflection within first-order logic has a great deal of standard machinery which we can then invoke; for more realistic agents, first-order logic is not a good representational fit to most real-world environments outside a human-constructed computer chip with thermodynamically expensive crisp variable states.

We will assume $\mathcal{A}^1$ is able to verify by inspection of $\mathcal{A}^0$'s hypothetical design[6] that (2.1) and (2.2) would be true about $\mathcal{A}^0$ if $\mathcal{A}^0$ were actually constructed:

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow (\forall b_i \in Acts^0 : \bar{b}_i \rightarrow \Box_0 \ulcorner \bar{b}_i \rightarrow \mathcal{G} \urcorner), \tag{2.3}$$

where $\bar{\mathcal{A}}^0$ denotes the proposition that $\mathcal{A}^0$ is constructed[7] and $\Box_0 \ulcorner \phi \urcorner$ denotes that there exists a proof of ϕ from Axm^0. Thus $\mathcal{A}^1$ proves that if $\mathcal{A}^0$ is constructed then the realization of any action b_i implies that $\mathcal{A}^0$ proved (from Axm^0) that $\bar{b}_i \rightarrow \mathcal{G}$.[8]

However for $\mathcal{A}^1$ to conclude that the goal $\mathcal{G}$ will actually be achieved by constructing $\mathcal{A}^0$, requires that $\mathcal{A}^1$ have axioms licensing the further reasoning step:

$$\mathcal{A}^1 \Vdash \forall b_i : (\Box_0 \ulcorner \bar{b}_i \rightarrow \mathcal{G} \urcorner \rightarrow (\bar{b}_i \rightarrow \mathcal{G})). \tag{2.4}$$

We can view this as a specified instance of a generic *soundness schema* over all formulas ϕ free in 0 or 1 variables:

$$\forall \phi : \mathcal{A}^1 \Vdash \forall x : \Box_0 \ulcorner \phi(x) \urcorner \rightarrow \phi(x). \tag{2.5}$$

For $\mathcal{A}^1$ to trust $\mathcal{A}^0$, the system Axm^1 must be able to prove that whenever a proof exists in Axm^0 of $\ulcorner \phi(x) \urcorner$, then $\phi(x)$ must be true. Löb's Theorem normally implies that no system can prove its own soundness schema—only strictly more powerful systems can prove a soundness schema over weaker systems.[9] This seemingly implies that $Axm^1 \neq Axm^0$ and that the offspring $\mathcal{A}^0$ must possess strictly weaker mathematical power than its parent $\mathcal{A}^1$, which is the Löbian obstacle. For purposes of completing the outline of an agent constructing another agent, we will assume for now that $\mathcal{A}^1$ is indeed of strictly greater mathematical power than $\mathcal{A}^0$ and continue.[10]

From (2.5), which yields (2.4) as a special case, combined with (2.3), $\mathcal{A}^1$ may conclude:

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow (\forall b_i \in Acts^0 : (\bar{b}_i \rightarrow \mathcal{G})). \tag{2.6}$$

$\mathcal{A}^1$ believes that if $\mathcal{A}^0$ takes any action, the goal $\mathcal{G}$ will be achieved. This follows from $\mathcal{A}^1$'s belief that if $\mathcal{A}^0$ takes any action then $\mathcal{A}^0$ will have proven that action to achieve $\mathcal{G}$, and from $\mathcal{A}^1$'s belief that $\mathcal{A}^0$'s reasoning is sound.

---

6. By Rice's Theorem no nontrivial property of output can be determined over arbitrary programs, but of course an AI, like a human engineer, may choose to work within a more restrictive space of programs which it can understand.

7. If the system were cashed out in greater detail, the proposition $\bar{\mathcal{A}}^0$ might appear as a giant conjunct of motor actions constructing the computer and program out of environmental transistors.

8. (2.3) is meant to represent a property which can be determined straightforwardly by inspection of code—e.g., by inspecting the code of $\mathcal{A}^0$, it is visible that no action b_i is ever sent to the motor output without passing through a check that syntactically verifies a proof of $\bar{b}_i \rightarrow \mathcal{G}$.

9. Suppose that $\mathcal{T}$ proves its own soundness schema $\forall \phi : \mathcal{T} \vdash \Box_{\mathcal{T}} \ulcorner \phi \urcorner \rightarrow \phi$. If $\phi = \bot$ then $\mathcal{T} \vdash \Box_{\mathcal{T}} \ulcorner \bot \urcorner \rightarrow \bot$ and $\mathcal{T}$ is inconsistent by Gödel's Second Incompleteness Theorem. See section 3.

10. E.g., suppose that $\mathcal{A}^1$ uses Zermelo-Fraenkel set theory ($\mathcal{ZF}$) as a base system while $\mathcal{A}^0$ uses Peano arithmetic. $\mathcal{ZF}$ can easily prove $\mathcal{PA}$ consistent, and sound with respect to $\mathcal{N}$ (the least set of ordinals closed under succession in that model of $\mathcal{ZF}$) in the sense of semantic entailment later described in section 6.

Suppose that $\mathcal{A}^0$ fails to take any action, e.g., because the agent $\mathcal{A}^0$ ran out of time before it could prove any theorem of the form $\bar{b}_i \rightarrow \mathcal{G}$? Let the special proposition $\emptyset$ denote a null or "default" outcome—the goal $\mathcal{G}$ has not been achieved, but neither has the environment been much disturbed (apart from some transistors having been configured into a useless but harmless agent $\mathcal{A}^0$).[11]

Allowing for $\emptyset$ to result when $\mathcal{A}^0$ fails to prove any theorem of the form $\bar{b}_i \rightarrow \mathcal{G}$, and letting $\bar{\mathcal{A}}^0_j$ denote the proposition that the construction of design $\mathcal{A}^0_j$ is implemented, $\mathcal{A}^1$'s criterion of action will be:

$$\bar{\mathcal{A}}^0_j \quad \Rightarrow \quad \mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0_j \rightarrow (\emptyset \vee \mathcal{G}). \tag{2.7}$$

For $\mathcal{A}^1$ to prove a theorem of the form (2.7) we need to make use of the further fact that if $\mathcal{A}^0$ does nothing, $\emptyset$ will result:

$$\bar{\mathcal{A}}^0 \quad \Rightarrow \quad (\forall b_i \in Acts^0 : \neg \bar{b}_i) \rightarrow \emptyset, \tag{2.8}$$

which we will suppose can be verified by $\mathcal{A}^1$ inspecting the hypothetical design[12] of $\mathcal{A}^0$:

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow ((\forall b_i : \neg \bar{b}_i) \rightarrow \emptyset). \tag{2.9}$$

Then from (2.9) and (2.6) $\mathcal{A}^1$ can obtain:

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow (\emptyset \vee \mathcal{G}), \tag{2.10}$$

which fulfills $\mathcal{A}^1$'s criterion of action (2.7) and permits $\mathcal{A}^0$ to be built.

Note that the specified architecture did not in fact tile from the design of $\mathcal{A}^0$ to the design of its ancestor $\mathcal{A}^1$—where $\mathcal{A}^0$ had the criterion of action $\bar{b}_i \Rightarrow \mathcal{A}^0 \Vdash \bar{b}_i \rightarrow \mathcal{G}$, its parent $\mathcal{A}^1$ needed the criterion of action $\bar{\mathcal{A}}^0_j \Rightarrow \mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0_j \rightarrow (\emptyset \vee \mathcal{G})$. However this new criterion of action will tile all aspects of the system—except the decreasing mathematical strength of each successive axiom system—without further change:

11. Suppose that the environment contains two registers, currently both set to 0, and the goal $\mathcal{G}$ is to set the registers to two numbers whose sum is 8 and product is 12. $\emptyset$ might then represent the set of outcomes where both registers remain set to 0—regardless of which environmental transistors have been configured or reconfigured, no agent has actually touched or changed the registers. Perhaps a product of 12 and a sum of 8 receives \$20 utility, a product and sum of 0 receives a reward of \$0, and any other product and sum receives a penalty of \$-1000. We would like to place into this environment an agent powerful enough to construct other agents, prove that if any agent touches the registers it will result in a reward of \$20, and know that otherwise the registers will remain untouched resulting in the default reward of \$0, with the penalty of \$-1000 not being risked—and permit the initial agent to construct other agents as well.

12. E.g., $\mathcal{A}^1$ looks over the transistors which will compose $\mathcal{A}^0$ and verifies by straightforward inspection that if the system never proves a theorem of the form $b_i \rightarrow \mathcal{G}$, the external motor circuitry will never activate.

$$
\begin{array}{rcl}
\bar{\mathcal{A}}^1_k & \Rightarrow & \mathcal{A}^2 \Vdash \bar{\mathcal{A}}^1_k \to (\emptyset \vee \mathcal{G}) \\
& & \mathcal{A}^2 \Vdash \bar{\mathcal{A}}^1 \to \forall \mathcal{A}^0_j \in Acts^1 : \bar{\mathcal{A}}^0_j \to \Box_1 \lceil \bar{\mathcal{A}}^0_j \to (\emptyset \vee \mathcal{G}) \rceil \\
& & \mathcal{A}^2 \Vdash \bar{\mathcal{A}}^1 \to (\forall \mathcal{A}^0_j \in Acts^1 : \neg \bar{\mathcal{A}}^0_j) \to \emptyset \\
\forall \phi : & & \mathcal{A}^2 \Vdash \forall x : (\Box_1 \lceil \phi(x) \rceil \to \phi(x)) \\
& & \mathcal{A}^2 \Vdash \bar{\mathcal{A}}^1 \to \forall \mathcal{A}^0_j \in Acts^1 : \bar{\mathcal{A}}^0_j \to (\emptyset \vee \mathcal{G}) \\
& & \mathcal{A}^2 \Vdash \bar{\mathcal{A}}^1 \to (\emptyset \vee \mathcal{G})
\end{array}
$$

□

3 The Löbian obstacle

Gödel's Second Incompleteness Theorem states that systems at least as powerful as Peano Arithmetic prove their own consistency if and only if they are inconsistent:

$$\mathcal{T} \vdash \neg \Box_{\mathcal{T}} \lceil \bot \rceil \iff \mathcal{T} \vdash \bot.$$

Löb's Theorem generalizes to prove that for any formula ϕ and any $\mathcal{T}$ at least as powerful as $\mathcal{PA}$:

$$\mathcal{T} \vdash \Box_{\mathcal{T}} \lceil \phi \rceil \to \phi \iff \mathcal{T} \vdash \phi.$$

Trivially, $\mathcal{T} \vdash \phi \Rightarrow \mathcal{T} \vdash \Psi \to \phi$, so the surprising statement is that a proof within $\mathcal{T}$ of $\Box_{\mathcal{T}} \lceil \phi \rceil \to \phi$ can be directly used to prove ϕ. With $\phi = \bot$ this yields (an intuitionistic proof of) the Second Incompleteness Theorem.

Gödel's sentence $G : \mathcal{PA} \vdash G \leftrightarrow \neg \Box_{\mathcal{PA}} \lceil G \rceil$ can be viewed as a non-paradoxical analogue of the Epimenides Paradox "This sentence is false." By a similar diagonalization over provability, Löb's Theorem constructs a Löb sentence $L : \mathcal{PA} \vdash L \leftrightarrow (\Box_{\mathcal{PA}} \lceil L \rceil \to \phi)$ which is a non-paradoxical analogue of the Santa Claus Paradox "If this sentence is true then Santa Claus exists." (Suppose the sentence were true. Then its antecedent would be true, the conditional would be true and thus Santa Claus would exist. But this is precisely what the sentence asserts, so it is true and Santa Claus does exist.)

The proof proceeds from the observation that $\mathcal{PA} \vdash L \to (\Box_{\mathcal{PA}} \lceil L \rceil \to \phi) \Rightarrow \mathcal{PA} \vdash \Box_{\mathcal{PA}} \lceil L \rceil \to \Box_{\mathcal{PA}} \lceil \phi \rceil$. From a model-theoretic standpoint, even when L has no standard proof, we would intuitively expect that every non-standard model of $\mathcal{PA}$ containing a nonstandard proof of L will also contain a nonstandard proof of ϕ; hence by the Completeness Theorem this should be

provable in $\mathcal{PA}$. Letting $\Box\ulcorner\Psi\urcorner \equiv \Box_{\mathcal{PA}}\ulcorner\Psi\urcorner$, the actual proof pathway is:

$$\begin{array}{ll}
\mathcal{PA} \vdash L \rightarrow (\Box\ulcorner L\urcorner \rightarrow \phi) & \\
\mathcal{PA} \vdash \Box\ulcorner L \rightarrow (\Box\ulcorner L\urcorner \rightarrow \phi)\urcorner & (\text{because } \mathcal{PA} \vdash \Psi \Rightarrow \mathcal{PA} \vdash \Box\ulcorner\Psi\urcorner) \\
\mathcal{PA} \vdash \Box\ulcorner L\urcorner \rightarrow \Box\ulcorner\Box\ulcorner L\urcorner \rightarrow \phi\urcorner & (\text{because } \mathcal{PA} \vdash \Box\ulcorner\Psi \rightarrow \Theta\urcorner \Rightarrow \\
 & \quad \mathcal{PA} \vdash \Box\ulcorner\Psi\urcorner \rightarrow \Box\ulcorner\Theta\urcorner) \\
\mathcal{PA} \vdash \Box\ulcorner L\urcorner \rightarrow (\Box\ulcorner\Box\ulcorner L\urcorner\urcorner \rightarrow \Box\ulcorner\phi\urcorner) & \\
\mathcal{PA} \vdash \Box\ulcorner L\urcorner \rightarrow \Box\ulcorner\Box\ulcorner L\urcorner\urcorner & (\text{because } \mathcal{PA} \vdash \Box\ulcorner\Psi\urcorner \rightarrow \Box\ulcorner\Box\ulcorner\Psi\urcorner\urcorner) \\
\mathcal{PA} \vdash \Box\ulcorner L\urcorner \rightarrow \Box\ulcorner\phi\urcorner &
\end{array}$$

Then if furthermore we have the Löbian hypothesis $\mathcal{PA} \vdash \Box\ulcorner\phi\urcorner \rightarrow \phi$ (a proof of ϕ implies that ϕ is true, or semantically, ϕ is true in every model where there exists a proof of ϕ):

$$\begin{array}{l}
\mathcal{PA} \vdash \Box\ulcorner\phi\urcorner \rightarrow \phi \\
\mathcal{PA} \vdash \Box\ulcorner L\urcorner \rightarrow \phi \\
\mathcal{PA} \vdash L \\
\mathcal{PA} \vdash \Box\ulcorner L\urcorner \\
\mathcal{PA} \vdash \phi.
\end{array}$$

Hence Löb's Theorem stating that in a theory $\mathcal{T}$ at least as powerful as $\mathcal{PA}$:

$$\mathcal{T} \vdash (\Box_{\mathcal{T}}\ulcorner\phi\urcorner \rightarrow \phi) \;\Rightarrow\; \mathcal{T} \vdash \phi.$$

Intuitively, $\mathcal{T}$ cannot trust its own proofs in the abstract: $\mathcal{T}$ cannot say, "I don't know if ϕ is true, but I trust that *if any proof exists* from my axioms of ϕ, then ϕ must be true." Instead $\mathcal{T}$ can only react to *particular* proofs of ϕ by asserting ϕ. This presents an apparent paradox of rational coherence when you try to interpret $\mathcal{T}$ as the beliefs of an agent; the agent will believe any proof of ϕ you present to it, but will refuse to assent to "I ought to believe any proof of ϕ I am presented with," a seeming defect of reflectivity—the agent cannot approve, under reflection, its own behavioral principles.

This presents an obstacle to perfectly tiling self-modification with $\mathcal{A}^1 \cong \mathcal{A}^0$ at the point where we would like $\mathcal{A}^1$ to carry out the inference:

$$\frac{\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow (\forall b_i \in Acts^0 : \bar{b}_i \rightarrow \Box_0\ulcorner\bar{b}_i \rightarrow (\emptyset \vee \mathcal{G})\urcorner)}{\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow (\forall b_i \in Acts^0 : \bar{b}_i \rightarrow (\emptyset \vee \mathcal{G}))}$$

If each ancestor $\mathcal{A}^{n+1}$ requires a successively more powerful proof system than its offspring $\mathcal{A}^n$ then this is analogous to a finite "telomere," which shortens in each constructed offspring (or shortens on each self-modification). In principle, this telomere can be made arbitrarily large, since from any base theory $\mathcal{T}$ we can construct more and more powerful theories by repeatedly adjoining soundness

schemas:

$$\begin{aligned}
\mathcal{T}{+}1 &\triangleq \mathcal{T} \cup \forall\phi: (\forall x : [\Box_{\mathcal{T}}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)]) \\
\mathcal{T}{+}2 &\triangleq \mathcal{T}{+}1 \cup \forall\phi: (\forall x : [\Box_{\mathcal{T}+1}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)]) \\
&... \\
\mathcal{T}{+}\omega &\triangleq \mathcal{T} \cup \forall\phi: (\forall n : \forall x : [\Box_{\mathcal{T}+n}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)]) \\
\mathcal{T}{+}\omega{+}1 &\triangleq \mathcal{T} \cup \forall\phi: (\forall x : [\Box_{\mathcal{T}+\omega}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)]) \\
&...
\end{aligned}$$

We can thus use the $\mathcal{T}{+}\alpha$ notation over any computable ordinal α, and indeed for any computable ordinals α, β the corresponding agent $\mathcal{A}^\alpha$ who believes in the theory $\mathcal{T}{+}\alpha$ can create the corresponding agent $\mathcal{A}^\beta$ iff $\beta < \alpha$. Thus by the well-ordering of the ordinals, such an agent $\mathcal{A}^\alpha$ can only create finite chains of descendants. E.g., the agent $\mathcal{A}^\omega$ trusting $\mathcal{T}{+}\omega$ must on its next step create an offspring trusting $\mathcal{T}{+}n$ for some finite n. Likewise all other well-founded systems of trust above $\mathcal{T}$ will reach the base system $\mathcal{T}$ after a finite descent.

That this is substantively decreasing the mathematical strength of the resulting agents can be illuminated by considering how $\mathcal{PA}{+}1$ is substantively stronger than $\mathcal{PA}$. Given any particular Goodstein sequence (Goodstein 1944), e.g., the sequence Goodstein(4) = 4, 26, 41, 60..., $\mathcal{PA}$ can prove that the sequence $G_1(4), G_2(4), G_3(4), ...$ will eventually reach 0. However, proving that $G(n)$ halts for larger and larger n requires $\mathcal{PA}$ to deploy proofs involving an increasing number of logical quantifiers $\forall x : \exists y : ...$ in its propositions. Thus $\mathcal{PA}$ cannot prove $\forall n : \exists x : G_x(n) = 0$ because this would require an infinite number of quantifiers to prove within $\mathcal{PA}$. A similar situation holds with respect to Kirby-Paris hydras, in which for any particular hydra, $\mathcal{PA}$ can prove that every strategy for cutting off that hydra's heads is a winning strategy, but as the hydra's heights increase so does the required number of quantifiers in the proof. Thus within $\mathcal{PA}$ it is not possible to prove that *every* hydra is killed by every strategy (Kirby and Paris 1982). In both cases the proofs have regular structure, so $\mathcal{PA}$ can describe how a proof of depth n can be formed for a hydra of height n, or $\mathcal{PA}$ can describe how to form a proof for the Goodstein sequence of any number, thus $\mathcal{PA}$ can prove:

$$\mathcal{PA} \vdash \forall n : \Box_{\mathcal{PA}}\ulcorner\exists x : G_x(n) = 0\urcorner.$$

But $\mathcal{PA}$ still cannot prove:

$$\mathcal{PA} \nvdash \forall n : \exists x : G_x(n) = 0.$$

This corresponds to what we earlier called a defect of reflective coherence, the state of believing "For every x, I believe x is true" but not believing "For every x, x is true." And thus $\mathcal{PA}{+}1$, augmented by the schema $\forall x : \Box_{\mathcal{PA}}\ulcorner\phi(x)\urcorner{\rightarrow}\phi(x)$, is able to prove that all Goodstein sequences halt and that all Kirby-Paris hydras are defeated.[13]

13. As of this very early draft, the above mathematical reasoning has not been verified. It looks obviously true to us that $\mathcal{PA}{+}1$ proves that all Goodstein sequences halt, but we still need to check.

More generally this state of affairs arises because the proof-theoretic ordinal of $\mathcal{PA}$ is ϵ_0, the limit of the ordinals $\omega, \omega^\omega, \omega^{\omega^\omega}, \ldots$.[14] Thus $\mathcal{PA}$ can prove the well-ordering of ordinal notations beneath ϵ_0,[15] but as Gentzen ([1936] 1969) showed, from the well-ordering of ϵ_0 itself it is possible to prove the syntactic consistency of $\mathcal{PA}$, and thus $\mathcal{PA}$ itself can never prove the well-ordering of an ordinal notation for ϵ_0.[16] Since the proof-theoretic ordinal of a mathematical system corresponds to its proof power in a deep sense, for each successive agent to believe in mathematics with a lower proof-theoretic ordinal would correspond to a substantive decrease in mathematical power.[17]

We are not the first to remark on how the inability of a theory to verify its own soundness schema can present apparent paradoxes of rational agency and

14. Expanding: 0 is the least ordinal. 1 is the first ordinal greater than 0. The first ordinal greater than all of $0, 1, 2, 3...$ is ω. The limit of $\omega, \omega{+}1, \omega{+}2, ...$ is 2ω. The limit of $\omega, 2\omega, 3\omega...$ is ω^2. The limit of $\omega, \omega^2, \omega^3, ...$ is ω^ω. The limit of $\omega^\omega, 2\omega^\omega, 3\omega^\omega, ...$ is $\omega^{\omega+1}$. Then ϵ_0 is the first ordinal greater than $\omega, \omega^\omega, \omega^{\omega^\omega}, ...$

15. An ordered pair (x, y) of natural numbers is a notation for the ordinals less than (but not equal to) ω^2, since in this ordering we first have $(0,0), (1,0), (2,0)$ for ω elements, followed by the ω elements for $(0,1), (1,1), (2,1)$ each of which is greater than all the preceding elements, and so on through ω copies of ω, but the notation does not contain any superelement $(0,0,1)$ which is the first element greater than all the preceding elements, so it does not contain a notation for ω^2. A notation ψ with a corresponding ordering $\psi_x < \psi_y$ can be shown to be a well-ordering if there are no infinite descending sequences $\psi_1 > \psi_2 > \psi_3...$, e.g., in the case above there is no infinite descending sequence $(2,2), (1,2), (0,2), (9999,1), ...$ even though the first number can jump arbitrarily each time the second number diminishes. For any particular ordinal $< \epsilon_0$, $\mathcal{PA}$ can show that a notation corresponding to that ordinal is well-ordered, but $\mathcal{PA}$ cannot show that any notation for *all* the ordinals less than ϵ_0 is well-ordered.

16. By assigning quoted proofs in $\mathcal{PA}$ to ordinals $< \epsilon_0$, it can be proven within $\mathcal{PA}$ that, if there exists a proof of a contradiction within $\mathcal{PA}$, there exists another proof of a contradiction with a lower ordinal under the ordering $<_{\epsilon_0}$. Then if it were possible to prove within $\mathcal{PA}$ that the ordering $<_{\epsilon_0}$ had no infinite descending sequences, $\mathcal{PA}$ would prove its own consistency. Similarly, $\mathcal{PA}$ can prove that any particular Goodstein sequence halts, but not prove that all Goodstein sequences halt, because the elements of any Goodstein sequence can be assigned to a decreasing series of ordinals $< \epsilon_0$. Thus any particular Goodstein sequence starts on some particular ordinal $< \epsilon_0$ and $\mathcal{PA}$ can prove that a corresponding notation is well-ordered and thence that the sequence terminates.

17. If you can show that the steps of a computer program correspond to a decreasing series of ordinals in some ordinal notation, you can prove that program will eventually halt. Suppose you start with a total (always-halting) computer program which adds 3, and you are considering a computer program which recursively computes 3^n via a function F of (x, y) with $F(0,1) = 0$, $F(x,1) = F(x-1,1) + 3$, and $F(x,y) = F(F(x,y-1),1)$ so that $F(1,n) = 3^n$. If you believe that the (x, y) notation is well-ordered then you can observe that each function call $F(\alpha) : \alpha \in (x, y)$ only calls itself with arguments $\beta < \alpha$, and hence that the corresponding tree of function calls must be finite. It is not uncommon for termination proofs in computer science to make use of ordinals much greater than ϵ_0, e.g., Kruskal's tree theorem (Kruskal 1960) or the strong normalization proof for System F (Girard 1971). Similarly by comprehending the well-ordering of notations for larger and larger ordinals, it is possible to prove the consistency of more and more powerful mathematical theories, e.g. $\mathcal{PA}$ corresponds to ϵ_0, Kripke-Platek set theory corresponds to the Bachmann-Howard ordinal, etc. Thus a mind losing its ability to recognize recursive notations as well-ordered, is indeed decreasing in substantive mathematical strength.

reflective coherence. Weaver (2005) made similar remarks centering on systems that can prove the well-ordering of any particular ordinal notation below the ordinal Γ_0 (representing the strength of systems mildly stronger than Peano arithmetic), and thus can prove that they prove notations for all ordinals below Γ_0, but cannot prove the combined ordinal notation for Γ_0 itself:

> Suppose A is a rational actor who has adopted some foundational stance. Any attempt to precisely characterize the limits of A's reasoning must meet the following objection: if we could show that A would accept every member of some set of statements S, then A should see this too and then be able to go beyond S . . . A can indeed see, as we do, that there exists a proof that he would accept for each statement in S, but he cannot go from this to actually accepting every statement in S . . . It is difficult to imagine a plausible set of beliefs that would not allow him to take this step.

Since proof-theoretic ordinals play such a basic role in describing the strength of mathematical theories—a theory which has been analyzed to have proof-theoretic ordinal α can usually prove the consistency of another theory with proof-theoretic ordinal β iff $\alpha > \beta$—it would seem reasonable for a pessimist to suspect that this form of "rational coherence" violates Gödel's Theorem and can never be obtained, and that we shall end up with agents who can only construct offspring who believe in math with lower proof-theoretic ordinals. It will be illuminating to consider the failure of some obvious-seeming attempts to bypass the Löbian obstacle:

(a) Via a standard diagonalization, in some theory $\mathcal{T}$ construct the fixed-point statement $\psi \leftrightarrow \phi(x) \vee \Box_{\mathcal{T}}\ulcorner\psi(x)\urcorner$. Then let $\phi(b_i) \equiv \bar{b}_i \rightarrow (\emptyset \vee \mathcal{G})$ (so that ψ then states, "Either x implies $(\emptyset \vee \mathcal{G})$, or you can prove ψ about x"). Then let $\mathcal{A}^\alpha$'s criterion of action be $\bar{b}_i \Rightarrow \mathcal{A}^\alpha \Vdash \psi(b_i)$.

Hope: $\mathcal{A}^\alpha$ reasoning about an agent $\mathcal{A}^\beta$ with $\bar{c_j} \Rightarrow \mathcal{A}^\beta \Vdash \psi(c_j)$ will be able to reason:

$$\mathcal{A}^\alpha \Vdash \bar{\mathcal{A}}^\beta \rightarrow \forall c_j : \bar{c_j} \rightarrow \Box_{\mathcal{T}}\ulcorner\psi(c_j)\urcorner$$
$$\mathcal{A}^\alpha \Vdash \bar{\mathcal{A}}^\beta \rightarrow \forall c_j : \bar{c_j} \rightarrow \psi(c_j)$$
$$\mathcal{A}^\alpha \Vdash \psi(\mathcal{A}^\beta)$$

Problem: $\Box_{\mathcal{T}}\ulcorner\psi(x)\urcorner \rightarrow \psi(x)$ so by Löb's Theorem $\forall x : \mathcal{T} \vdash \psi(x)$.[18]

(b) Instead of an indefinitely extensible ordinal hierarchy of systems $\mathcal{T}+\alpha$, let us form a descending sequence of theories $\tilde{\mathcal{T}}$–0, $\tilde{\mathcal{T}}$–1, $\tilde{\mathcal{T}}$–2 as follows:

$$\tilde{\mathcal{T}}\text{–}n \triangleq \mathcal{T} \cup \forall\phi : (\forall x : \Box_{\tilde{\mathcal{T}}\text{–}(n+1)}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)).$$

Hope: Each theory n trusts all lower theories $> n$, but no theory trusts itself. Thus the corresponding agent $\mathcal{A}^\alpha$ which starts at $\tilde{\mathcal{T}}$–0 will trust the

18. $\psi(x)$ is just a Henkin sentence $H \leftrightarrow \Box_{\mathcal{T}}\ulcorner H\urcorner$ with a dangling $\vee$ clause $\phi(x)$. A Henkin sentence for $\mathcal{T}$ is of course always provable within $\mathcal{T}$.

reasoning of an agent $\mathcal{A}^\beta$ which uses $\tilde{\mathcal{T}}$–1, and so on tiling indefinitely. The "problem of immediate reflection" (on considering one's own source code) will remain, but it will at least be possible to carry out an indefinitely extending sequence of agent constructions without the strength of trusted mathematics ever decreasing below the base $\mathcal{T}$.

Problem: If $\mathcal{T}$ is at least as strong as $\mathcal{PA}$ it will be straightforward to prove within $\tilde{\mathcal{T}}$–0 that any proof in $\tilde{\mathcal{T}}$–0 can be translated into a proof in $\tilde{\mathcal{T}}$–1 (by increasing all the $\tilde{\mathcal{T}}$–x numbers x by 1 inside the proof). Thus:

$$\tilde{\mathcal{T}}\text{–}0 \vdash \Box_{\tilde{\mathcal{T}}-0}\ulcorner\bot\urcorner \rightarrow \Box_{\tilde{\mathcal{T}}-1}\ulcorner\bot\urcorner$$
$$\tilde{\mathcal{T}}\text{–}0 \vdash \Box_{\tilde{\mathcal{T}}-1}\ulcorner\bot\urcorner \rightarrow \bot$$
$$\tilde{\mathcal{T}}\text{–}0 \vdash \Box_{\tilde{\mathcal{T}}-0}\ulcorner\bot\urcorner \rightarrow \bot$$
$$\tilde{\mathcal{T}}\text{–}0 \vdash \bot$$

(c) Perhaps the previous constructions failed due to their impredicativity. Let us try a finitary version which speaks only of well-founded objects. Let $\Box_{\mathcal{T}}^{k+1}\ulcorner\phi\urcorner$ stand for the formula $\Box_{\mathcal{T}}\ulcorner\Box_{\mathcal{T}}\ulcorner...\phi...\urcorner\urcorner$ with $k+1$ boxes.. Since the pattern $\Box_{\mathcal{T}}\ulcorner\Box_{\mathcal{T}}\ulcorner...\urcorner\urcorner$ is highly regular, a compact formula for $\Box_{\mathcal{T}}^{k+1}\ulcorner\phi\urcorner$ should exist. Then let $\mathcal{A}^\alpha$'s criterion of action be $\bar{b}_i \Rightarrow \mathcal{A}^\alpha \Vdash \phi(b_i) \vee \Box_{\mathcal{T}}^{k+1}\ulcorner\phi(b_i)\urcorner$ where $\phi(b_i) \equiv \bar{b}_i \rightarrow (\emptyset \vee \mathcal{G})$ as before. Let $\mathcal{A}^\beta$'s criterion of action be the same.

Hope: Rather than trying to modify $\mathcal{A}^\alpha$'s epistemic beliefs in a way that will circumvent Löb's Theorem, we are relaxing $\mathcal{A}^\alpha$'s instrumental decision criterion to permit $\mathcal{A}^\alpha$ to do b_i whenever b_i leads to $\mathcal{G}$, provably leads to $\mathcal{G}$, provably provably leads to $\mathcal{G}$, etc. There is no Gödelian prohibition against a more permissive decision criterion; and from outside the system, if we trust the soundness of $\mathcal{T}$+1, we will trust that in all such cases $\mathcal{G}$ will be achieved. $\mathcal{A}^\alpha$ inspecting $\mathcal{A}^\beta$ will conclude that whenever $\mathcal{A}^\beta$ does c_j, there exists a statement with one or more levels of proof-ness that $\bar{c_j} \rightarrow (\emptyset \vee \mathcal{G})$ and thence conclude $(\mathcal{A}^\beta \rightarrow \emptyset) \vee \Box_{\mathcal{T}}^{k+1}\ulcorner\mathcal{A}^\beta \rightarrow (\emptyset \vee \mathcal{G})\urcorner$. Furthermore, because of the *perfect* tiling, $\mathcal{A}^\alpha$ will trust its own source code upon reflecting on it (since this merely adds another level of proof-ness). Furthermore, $\mathcal{A}^\alpha$ will behave with perfect naturalism in the sense that any environmental computing system believed to implement $\mathcal{T}$ will provide just as much justification-for-action as any internal proof from $\mathcal{T}$'s axioms; there will be no non-naturalistic discontinuity between observations of outputs of $\mathcal{A}^\alpha$'s internal subroutines, and observations of the outputs of environmental systems which $\mathcal{A}^\alpha$ believes to have isomorphic properties to its beliefs about its internal subroutines.

Problem: The formula $\Box_{\mathcal{T}}^{k+1}\ulcorner\phi(b_i)\urcorner$ must actually be written as $\exists k : \Box_{\mathcal{T}}^{k+1}\ulcorner\phi(b_i)\urcorner$. Then when $\mathcal{A}^\alpha$ considers $\mathcal{A}^\beta$ taking an action c_j, $\mathcal{A}^\alpha$ can conclude only that it must have been the case that $\Box_{\mathcal{T}}\ulcorner\exists k : \Box_{\mathcal{T}}^{k+1}\ulcorner\phi(b_i)\urcorner \vee \phi(b_i)\urcorner$, which does not reduce to $\exists k : \Box_{\mathcal{T}}^{k+1}\ulcorner\phi(b_i)\urcorner$.[19]

19. The formula $\Box_{\mathcal{T}}^{k+1}\ulcorner\phi\urcorner$ just occupies a new ordinal ω of proofness, and $\Box_{\mathcal{T}}\ulcorner\Box_{\mathcal{T}}^{k+1}\ulcorner\phi\urcorner\urcorner$ then corresponds to a new ordinal ω+1 of proofness.

At this point the mathematician already accustomed to seeing such systems stratify themselves, or else collapse into inconsistency, may be wondering whether it should even be normatively possible to bypass the Löbian obstacle. Yet it seems intuitively like it ought to be possible to have an agent that "acts as if" Zermelo-Fraenkel set theory is true, including constructing other agents which "act as if" $\mathcal{ZF}$; this does not seem like an instrumentally incoherent way for a rational agent to behave. We have presented the failed hopes above in order to better illustrate some of the further desiderata of self-modifying agents which we could ask from a solution to the Löbian obstacle:

Indefinitely tiling trust: For any system $\mathcal{T}$ (e.g. $\mathcal{ZF}$) it should be possible to characterize an infinitely descending sequence of agents constructing agents (equivalently an indefinitely self-modifying agent) whose level of trusted mathematics never drops below $\mathcal{T}$.

Reflectively coherent quantified belief: The agent will not encounter any context in which it believes $\forall x : \Box_{\text{agent}} \ulcorner \phi(x) \urcorner$ yet cannot bring itself to believe $\forall x : \phi(x)$.

Reflective trust: An agent reflecting on its beliefs about the program P composing itself, should not find that it distrusts P's actions or that it rejects P as an acceptable program; the agent's beliefs about its own construction should seem to license the beliefs it possesses, or at least license the performance of the actions which P has selected.[20]

Naturalistic trust: If an agent believes that an environmental system isomorphic to one of its own subroutines has yielded an output, this belief should license the same sort of actions as would be licensed by the corresponding output by the corresponding internal subroutine. If you are "acting as if" $\mathcal{PA}$ is true, then seeing an environmental theorem-prover that you believe to implement $\mathcal{PA}$ flashing "Proven: $2 + 2 = 4$!" should yield the same sort of epistemic updates and instrumental decisions as when your internal $\mathcal{PA}$-prover signals that "$2 + 2 = 4$" is a theorem.

At present we can partially resolve the Löbian obstacle by technical means to the extent of obtaining the first and second desiderata, but not yet the third and fourth desiderata, nor yet by fundamental rather than technical means.

4 Partial solutions to the Löbian obstacle

4.1 Infinitely descending soundness schemas

The fatal flaw in the previously presented sequence of theories $\tilde{\mathcal{T}}$–0, $\tilde{\mathcal{T}}$–1, $\tilde{\mathcal{T}}$–2... with $\tilde{\mathcal{T}}$–$n \triangleq \mathcal{T} \cup \forall \phi : (\forall x : \Box_{\tilde{\mathcal{T}}\text{–}(n+1)} \ulcorner \phi(x) \urcorner \rightarrow \phi(x))$ is that any proof in $\tilde{\mathcal{T}}$–0 provably mapped onto a proof in $\tilde{\mathcal{T}}$–1 and hence $\tilde{\mathcal{T}}\text{–}0 \vdash \Box_{\tilde{\mathcal{T}}\text{–}0} \ulcorner \bot \urcorner \rightarrow \Box_{\tilde{\mathcal{T}}\text{–}1} \ulcorner \bot \urcorner$.

20. In other words: A self-modifying AI at the level of $\mathcal{PA}$+2 which obtains a glimpse at its own source code should not immediately panic and self-modify to only believe theorems at the level of $\mathcal{PA}$+1, followed by an additional panic bringing it down to the level of $\mathcal{PA}$...

This flaw is repairable. Given a base theory $\mathcal{T}$, let $\psi(0), \psi(1), \psi(2)...$ be a sequence of statements such that $\mathcal{T} \vdash \psi(0), \mathcal{T} \vdash \psi(1), ...$ but $\mathcal{T} \nvdash (\forall n : \psi(n))$. E.g., suppose the theory $\mathcal{ZF}$ is much stronger than $\mathcal{T}$ (we shall soon define what sense of "much stronger" is required; if $\mathcal{T} = \mathcal{PA}$ then $\mathcal{ZF}$ will be much stronger in the appropriate sense). Then $\psi(n)$ can state "n is not the Gödel number of a proof of a contradiction in $\mathcal{ZF}$ ":

$$\psi(n) \equiv \neg Bew_{\mathcal{ZF}}(n, \ulcorner \bot \urcorner). \tag{4.1}$$

Since $Bew_{\mathcal{ZF}}$ is a Δ_0 formula, the truth of any particular $\psi(n)$ is provable in any $\mathcal{T}$ as strong as $\mathcal{PA}$, but proving $\forall n : \psi(n)$ would require $\mathcal{T} = \mathcal{PA}$ proving the consistency of Zermelo-Fraenkel set theory, which should not happen. Then let:

$$\mathcal{T}\text{–}n \triangleq \mathcal{T} \cup \forall \phi : \psi(n) \rightarrow (\forall x : \Box_{\mathcal{T}\text{–}(n+1)} \ulcorner \phi(x) \urcorner \rightarrow \phi(x)). \tag{4.2}$$

Since every theory holds $\psi(n)$ as a precondition for trusting the next lower theory in the infinite descending sequence, and the base theory $\mathcal{T}$ cannot prove $\forall n : \psi(n)$, it will be impossible for the base theory $\mathcal{T}$ to prove that a proof in $\mathcal{T}$–n always translates into a proof in $\mathcal{T}$–$(n{+}1)$.[21] We shall now show that $\mathcal{T}$–0 is consistent, by proving that $(\mathcal{T}\text{–}0 \vdash \bot) \rightarrow (\forall n : \psi(n))$ within a sound theory $\mathcal{T}^+$ which is too weak to prove $\forall n : \psi(n)$:

Without loss of generality let $\mathcal{T}$ be cast in a form where each step of a $\mathcal{T}$-proof is either an axiom or a modus ponens inference step of the form $\Gamma, \Gamma \rightarrow A \vdash A$. Let $TRUE_{\mathcal{T}}$ be a truth predicate over formulas in the language of $\mathcal{T}$.[22] Let $Ax(TRUE_{\mathcal{T}})$ be a set of axioms for standard reasoning about the syntax of the quoted arguments to $TRUE_{\mathcal{T}}$, including $\forall \ulcorner \Gamma \urcorner : \forall \ulcorner \Delta \urcorner : TRUE_{\mathcal{T}} \ulcorner \Gamma \rightarrow \Delta \urcorner \rightarrow (\neg TRUE_{\mathcal{T}} \ulcorner \Gamma \urcorner \vee TRUE_{\mathcal{T}} \ulcorner \Delta \urcorner)$ and $(\forall n : TRUE_{\mathcal{T}} \ulcorner \phi(n) \urcorner) \leftrightarrow TRUE_{\mathcal{T}} \ulcorner \forall n : \phi(n) \urcorner$. Then let T^+ equal the base theory $\mathcal{T}$ (in this case $\mathcal{PA}$) augmented by the truth predicate $TRUE_{\mathcal{T}}$, the soundness of $\mathcal{T}$, and the axioms $Ax(TRUE_{\mathcal{T}})$:

$$\begin{aligned} \mathcal{T}^+ \triangleq \mathcal{T} &\cup \forall \phi : \forall x : TRUE_{\mathcal{T}} \ulcorner \phi(x) \urcorner \leftrightarrow \phi(x) \\ &\cup \forall \ulcorner \phi \urcorner : \Box_{\mathcal{T}} \ulcorner \phi \urcorner \rightarrow TRUE_{\mathcal{T}} \ulcorner \phi \urcorner \\ &\cup Ax(TRUE_{\mathcal{T}}) \end{aligned} \tag{4.3}$$

(Note that $\mathcal{T}^+$ should be sound on the standard numbers $\mathcal{N}$ because every axiom of $\mathcal{T}^+$ is true within $\mathcal{N}$, and that if $\mathcal{T} = \mathcal{PA}$ then $\mathcal{T}^+$ should be too weak to prove $Con(\mathcal{ZF})$.)

The following reasoning can be carried out within $\mathcal{T}^+$. Suppose $\Box_{\mathcal{T}\text{–}0} \ulcorner \bot \urcorner$. Then there exists some proof p with $Bew_{\mathcal{T}\text{–}0}(p, \bot)$. Consider the 'first bad line'

21. Consider a sequence of theories $\dot{\mathcal{T}}$–n where $\mathcal{T} = \mathcal{PA}$ and where $\dot{\psi}(n)$ is false beyond some upper bound m, e.g. $\dot{\psi}(n) \iff n < m$. Then the waterfall of trust descends only a finite length before reaching bottom, and the induction argument for $\dot{\mathcal{T}}\text{–}n \ulcorner \phi \urcorner \rightarrow \dot{\mathcal{T}}\text{–}(n{+}1) \ulcorner \phi \urcorner$ fails because proofs in $\dot{\mathcal{T}}\text{–}m{-}1$ cannot translate into $\dot{\mathcal{T}}$–m. Indeed, $\dot{\mathcal{T}}$–0 is just equivalent to $\mathcal{PA}{+}m$, $\dot{\mathcal{T}}$–1 is equivalent to $\mathcal{PA}{+}(m{-}1)$, etc. with $\dot{\mathcal{T}}\text{–}m^+ \cong \mathcal{PA}$. Since $\mathcal{T}$ cannot prove $\forall n : \psi(n)$, it cannot prove the waterfall is not finite, and hence cannot prove the induction argument on $\dot{\mathcal{T}}\text{–}n \ulcorner \phi \urcorner \rightarrow \dot{\mathcal{T}}\text{–}(n{+}1) \ulcorner \phi \urcorner$.

22. In accordance with Tarski's theorem, the predicate $TRUE_{\mathcal{T}}$ can only talk about formulas in the base theory $\mathcal{T}$ and not in the augmented theory $\mathcal{T} + TRUE_{\mathcal{T}}$.

of p, the first step of p asserting Γ where $\neg TRUE_{\mathcal{T}}\ulcorner\Gamma\urcorner$. The first possibility is that this line is an axiom of the base theory $\mathcal{T}$ but all statements provable in $\mathcal{T}$ are $TRUE_{\mathcal{T}}$. The second possibility is that this line is an inference of the form $\Gamma, (\Gamma \rightarrow A) \vdash A$ but for this line to be $\neg TRUE_{\mathcal{T}}$ and all earlier lines to be $TRUE_{\mathcal{T}}$ requires $TRUE_{\mathcal{T}}\ulcorner\Gamma\urcorner \wedge TRUE_{\mathcal{T}}\ulcorner\Gamma \rightarrow A\urcorner \wedge \neg TRUE_{\mathcal{T}}\ulcorner A\urcorner$ which contradicts $Ax(TRUE_{\mathcal{T}})$. Thus the first bad line must invoke one of the special axioms in $\mathcal{T}$–0 of the form $\psi(0) \rightarrow (\forall x : \Box_{\mathcal{T}-1}\ulcorner\phi(x)\urcorner \rightarrow \phi(x))$. If this axiom is $\neg TRUE_{\mathcal{T}}$ then $\psi(0)$ must be $TRUE_{\mathcal{T}}$ and there must exist some formula $\ulcorner\phi\urcorner$ and number x such that $\Box_{\mathcal{T}-1}\ulcorner\phi(x)\urcorner$ is $TRUE_{\mathcal{T}}$ and ϕ (x) is $\neg TRUE_{\mathcal{T}}$. Thus there must exist a bad proof in $\mathcal{T}$–1 and this bad proof must also have a first bad line. By induction there must exist bad proofs in every $\mathcal{T}$–n and $\psi(n)$ must be true for all n. Therefore,

$$\mathcal{T}^{+} \vdash (\Box_{\mathcal{T}-0}\ulcorner\bot\urcorner) \rightarrow (\forall n : \psi(n)). \tag{4.4}$$

If there exists a proof of a contradiction within $\mathcal{T}$–0, this fact will be provable in $\mathcal{T}^{+}$ by exhibiting the proof. But $\mathcal{T}^{+}$ is sound and too weak to prove $\forall n : \psi(n)$. Therefore $\mathcal{T}$–0 is consistent. $\Box$

The corresponding agents $\mathcal{A}^{\alpha}$ are then constructed in the obvious fashion.

Remarks. Although $\mathcal{T}$–0 is slightly more powerful than $\mathcal{T}$–1 in the sense that $\mathcal{T}$–0 can prove certain exact theorems which $\mathcal{T}$–1 cannot, the proof-theoretic ordinal of every $\mathcal{T}$–n should be the same, and equal to the limit of the proof-theoretic ordinals of $\mathcal{T}$, $\mathcal{T}$+1, $\mathcal{T}$+2, ... and less than the proof-theoretic ordinal of $\mathcal{T}+\omega$, since any $\mathcal{T}$–n can invoke any finite number of layers of a soundness schema over $\mathcal{T}$, but cannot invoke quantified trust in an infinite number of layers as does $\mathcal{T}+\omega$. We thus answer in the affirmative "Can a consistent theory verify the soundness of another theory with the same proof-theoretic ordinal?" and "Can there be an indefinitely tiling sequence of agents whose trust never falls below a base theory?"

$\mathcal{T}$–n also provides a solution to the challenge of rational coherence posed by Weaver (2005), what we termed the desideratum of reflectively coherent quantified belief: An agent constructed around $\mathcal{T}$–n will not find itself saying "For every n, I believe ϕ (n) is true" when it cannot say "I believe that for every n, ϕ (n) is true." Through longer and longer finite proofs, $\mathcal{T}$–0 can prove the well-ordering of any ordinal notation provable in $\mathcal{T}$, $\mathcal{T}$+1, $\mathcal{T}$+2 ... but $\mathcal{T}$–0 does not know that it proves every ordinal notation in this series - $\mathcal{T}$–0 must observe the proof to know what it proves.

This is not to say that $\mathcal{T}$–0 is a reasonable representation of a rational agent's state of mind. However, besides offering a constructive example of an agent which technically meets Weaver's desideratum, it suggests that a coherent rational agent might be able to verify increasingly recursive ordinal notations via increasing amounts of thought, but never know the limit of what it will accept; and this agent may be able to construct offspring that likewise verify increasing ordinal notations in an equally powerful series and likewise do not know the limit of what they will accept.[23]

23. Which is, in general, the agenda of this paper: Our framework and our technical solutions are

Disadvantage. Since $\mathcal{T}$–0 is consistent it must have a model, and since all axioms of $\mathcal{T}$ are axioms of $\mathcal{T}$–0 this model must also be a model of $\mathcal{T}$. However we may still worry that, e.g., if $\mathcal{T}$ is $\mathcal{PA}$ then $\mathcal{T}$–0 may have only nonstandard models of arithmetic; perhaps $\mathcal{T}$–0 is not sound on the standard numbers $\mathcal{N}$. This fear is well-founded and in particular $\mathcal{T}$–$0 \vdash \exists n : \neg\psi(n)$ via:

$$\begin{aligned}
&\mathcal{T}\text{–}0 \vdash (\forall n : \psi(n)) \rightarrow \Box_{\mathcal{T}\text{–}0}\ulcorner(\forall n : \psi(n))\rightarrow\bot\urcorner \rightarrow \Box_{\mathcal{T}\text{–}1}\ulcorner(\forall n : \psi(n))\rightarrow\bot\urcorner \\
&\mathcal{T}\text{–}0 \vdash (\forall n : \psi(n)) \rightarrow \Box_{\mathcal{T}\text{–}0}\ulcorner(\forall n : \psi(n))\rightarrow\bot\urcorner \rightarrow ((\forall n : \psi(n))\rightarrow\bot) \\
&\mathcal{T}\text{–}0 \vdash \Box_{\mathcal{T}\text{–}0}\ulcorner(\forall n : \psi(n))\rightarrow\bot\urcorner \rightarrow ((\forall n : \psi(n))\rightarrow\bot) \\
&\mathcal{T}\text{–}0 \vdash (\forall n : \psi(n))\rightarrow\bot
\end{aligned} \tag{4.5}$$

One might perhaps argue that a belief that $\mathcal{ZF}$ is inconsistent is not too troubling, in the sense that any physical situation in which this belief gets an agent into trouble ought to correspond to a physical situation that demonstrates $\mathcal{ZF}$ to be consistent. Nonetheless we would like our agents to be able to have beliefs with a standard model. Otherwise the agent will falsely believe that a Turing machine seeking a proof of contradiction in $\mathcal{ZF}$ will halt; and this false belief further implies that the agent falsely believes that its sequence of offspring will inevitably come to a halt after some unknown finite time. This seems sufficient to exclude the $\mathcal{T}$–n family from direct consideration as the basis of a sufficiently advanced self-modifying agent.

4.2 Parametric polymorphism

Let $\mathcal{T}$ be a theory with models including the standard numbers $\mathcal{N}$, that is $\mathcal{N} \models \mathcal{T}$ ($\mathcal{N}$ semantically entails $\mathcal{T}$). Benja Fallenstein's "parametric polymorphism approach"[24] augments the language of $\mathcal{T}$ with an extra term κ which, from outside the system, is intended to refer to any natural number in $\mathcal{N}$. T_κ then contains a self-referential axiom schema asserting that if a statement ϕ is provable in T_κ and the constant κ happens to be greater than 0, then ϕ with all instances of κ replaced by κ–1 is true:

$$T_\kappa \triangleq \mathcal{T} \cup \forall\phi : (\kappa > 0) \rightarrow \forall x : \Box_{T_\kappa}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)[\kappa\backslash\kappa\text{–}1]. \tag{4.6}$$

not meant to be suitable, as they stand, to rational agents / sufficiently advanced self-modifying machine intelligences, which would e.g. presumably be probabilistic boundedly-rational agents rather than theorem provers, etc. The idea is rather to expose foreseeable difficulties of reflection for self-modifying agents and to some extent offer constructive demonstrations that these difficulties are solvable, even if the solution is technical rather than fundamental, thus hopefully moving us toward an eventually satisfactory theory of the reflectively coherent, boundedly-rational probabilistic agent, even if that later theory is constructed along quite different lines than the $\mathcal{A}^\alpha$ schema, as it almost certainly must be.

24. This workaround for the Löbian obstacle was initially developed by Benja Fallenstein (independently of Herreshoff's infinite descent above) in response to the informal challenge posed in Yudkowsky's conference presentation of Yudkowsky (2011), and initially posted to Fallenstein (2012). It was further adapted to the $\mathcal{A}^\alpha$ formalism shown here during Fallenstein's attendance at the April 2013 MIRI Workshop on Logic, Probability, and Reflection with some commentary by other workshop attendees. For the origin of the term "parametric polymorphism" see Strachey (1967).

We shall prove that T_κ is sound when κ is interpreted as any number in $\mathcal{N}$, and then present an infinite sequence of agents which prove their offspring "safe for κ steps". Since κ can be interpreted as any number, from outside the system we conclude that such agents are safe for any number of steps.

The proof is by induction on models $\{\mathcal{N}, \kappa{=}n\}$ of T_κ. For the base case $\{\mathcal{N}, \kappa{=}0\} \models T_\kappa$ observe that if $\kappa{=}0$ then the antecedent of every extra axiom is false and so the extra axiom schema is trivially true.

For the induction step assume $\{\mathcal{N}, \kappa{=}n\} \models T_\kappa$. Using this assumption we shall demonstrate in three steps that, for every $x \in \mathcal{N}$:

$$\{\mathcal{N}, \kappa{=}n{+}1\} \models \Box_{T_\kappa}\ulcorner\phi(x)\urcorner \Rightarrow \{\mathcal{N}, \kappa{=}n\} \models \Box_{T_\kappa}\ulcorner\phi(x)\urcorner \tag{4.7}$$

$$\{\mathcal{N}, \kappa{=}n\} \models \Box_{T_\kappa}\ulcorner\phi(x)\urcorner \Rightarrow \{\mathcal{N}, \kappa{=}n\} \models \phi(x) \tag{4.8}$$

$$\{\mathcal{N}, \kappa{=}n\} \models \phi(x) \Rightarrow \{\mathcal{N}, \kappa{=}n{+}1\} \models \phi(x)[\kappa\backslash\kappa{-}1] \tag{4.9}$$

(4.9) is obvious. For (4.7) observe that the set of proofs in T_κ and the values of $\Box_{T_\kappa}$ are unchanged from $\{\mathcal{N}, \kappa{=}n\}$ to $\{\mathcal{N}, \kappa{=}n{+}1\}$ since the provability formula Bew_{T_κ} makes no mention of κ.[25] Thus $\{\mathcal{N}, \kappa{=}n{+}1\} \models \Box_{T_\kappa}\ulcorner\psi\urcorner \Rightarrow \{\mathcal{N}, \kappa{=}n\} \models \Box_{T_\kappa}\ulcorner\psi\urcorner$ and we can simply write $\mathcal{N} \models \Box_{T_\kappa}\ulcorner\psi\urcorner$. To extend to formulas free in one variable $\phi(x)$, observe that since we are dealing with a natural model $\mathcal{N}$ of T_κ, for every object x there exists a corresponding finite term c of the language ($SSS...0$ with x repetitions of the successor function) and thus our proof extends via $\psi \equiv \phi(c)$.

For (4.8) observe that since we are dealing with a natural model $\mathcal{N}$ of T_κ, only natural proofs exist and the induction step $\{\mathcal{N}, \kappa{=}n\} \models T_\kappa$ gives us directly that $(\mathcal{N} \models \Box_{T_\kappa}\ulcorner\psi\urcorner) \Rightarrow (\{\mathcal{N}, \kappa{=}n\} \models \psi)$.[26] We extend over formulas free in one variable as above: Since the domain is the natural numbers $\mathcal{N}$, for every object x in the universe there will exist a term of the language c such that $(\psi \equiv \phi(c)) \iff \phi(x)$.

Therefore $\{\mathcal{N}, \kappa{=}n{+}1\} \models \forall x : \Box_{T_\kappa}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)$. This completes the induction step and thus:

$$\forall n : \{\mathcal{N}, \kappa{=}n\} \models T_\kappa \qquad \Box \tag{4.10}$$

The corresponding agents must be modified to represent the idea of "safe for a number of steps." We will introduce a time t on which any given action

25. Even in models where $\kappa > 0$, there is no proof that $\kappa > 0$, so the class of proofs which goes through is unchanged. The formula $Bew_{T_\kappa}(p, \ulcorner\psi\urcorner)$ will contain only formulas dealing with the Gödel-numbers of proofs in T_κ, and these Gödel-numbers themselves will contain no instances of κ, so which standard numbers are the Gödel numbers of T_κ-proofs does not change from model to model.

26. We are not obtaining $\{\mathcal{N}, \kappa{=}n\} \models (\Box_{T_\kappa}\ulcorner\psi\urcorner \rightarrow \psi)$ from the assumption that the axiom schema was true in the previous n, because perhaps $\kappa = 0$ in the previous n and the axiom schema was vacuously true. Rather since the model is $\mathcal{N}$ and there are no nonstandard proofs, any time there exists a Gödel number of a proof of an implication in classical logic $\Gamma \rightarrow \Delta$, it will be true across all models m that $m \models \Gamma \Rightarrow m \models \Delta$. By the induction postulate we have directly that $\{\mathcal{N}, \kappa{=}n\}$ is a model of T_κ. Then any statement ψ syntactically derivable from T_κ by a standard proof in classical logic, will also have $\{\mathcal{N}, \kappa{=}n\}$ as a model.

occurs and assume that if $\mathcal{A}^1$ constructs $\mathcal{A}^0$ at t then $\mathcal{A}^0$ will not act until at least time $t{+}1$ or later, and that this is apparent to $\mathcal{A}^1$ from inspection of $\mathcal{A}^0$.[27] Thus if $\mathcal{A}^0$ acts at time $t{+}1$ or later, and takes actions "safe" for an additional n ticks beyond that until $(t{+}1){+}n$, from $\mathcal{A}^1$'s perspective constructing $\mathcal{A}^0$ at time t is safe until $t{+}(n{+}1)$ ticks.

Let $\bar{b_{i,t}}$ be the proposition that b_i is done at time t, let $\bar{\mathcal{A}_t^0}$ be the proposition that $\mathcal{A}^0$ is built at t, let $\emptyset_t$ be the proposition that $\emptyset$ is true at t, and let $\mathcal{G}_t$ be the proposition that the goal has been achieved at t. If no action is taken before t then $\emptyset_t$. When quantifying over acts and times we shall implicitly demand $b_i \in Acts^0$, $\mathcal{A}_i^0 \in Acts^1$ and $t \in Time$. Then with T_κ as a base theory $\mathcal{A}^1$'s reasoning proceeds:

$$\bar{\mathcal{A}_t^0} \Rightarrow \mathcal{A}^1 \Vdash \bar{\mathcal{A}_t^0} \rightarrow (\forall u : u \leq t + \kappa \rightarrow (\emptyset_u \vee \mathcal{G}_u)) \qquad (\mathcal{A}^1 \text{ decisions}) \quad (4.11)$$

$$\bar{b_{i,t}} \Rightarrow \mathcal{A}^0 \Vdash \bar{b_{i,t}} \rightarrow (\forall u : u \leq t + \kappa \rightarrow (\emptyset_u \vee \mathcal{G}_u)) \qquad (\mathcal{A}^0 \text{ decisions}) \quad (4.12)$$

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}_t^0} t \rightarrow (\forall b_i : \forall v : \bar{b_{i,v}} \rightarrow \Box_0 \lceil \bar{b_{i,v}} \rightarrow \forall u : u \leq v + \kappa \rightarrow (\emptyset_u \vee \mathcal{G}_u) \rceil) \qquad (\mathcal{A}^1 \text{ knows 4.12}) \quad (4.13)$$

$$\mathcal{A}^1 \Vdash (\kappa > 0) \rightarrow (\forall b_i : \forall t : [\Box_0 \lceil \rightarrow \forall u \leq t + \kappa : (\emptyset_u \vee \mathcal{G}_u) \rceil \rightarrow (\rightarrow \forall v \leq t + \kappa - 1 : (\emptyset_v \vee \mathcal{G}_v))]) \qquad (\text{soundness schema for } T_\kappa) \quad (4.14)$$

$$\forall t : (\forall b_i : \forall u \leq t : \neg \bar{b_{i,u}}) \Rightarrow (\forall u \leq t : \emptyset_u) \qquad (\text{No action} \Rightarrow \emptyset_t) \quad (4.15)$$

$$\mathcal{A}^1 \Vdash \forall t : ((\forall b_i : \forall u \leq t : \neg \bar{b_{i,u}}) \rightarrow (\forall u \leq t : \emptyset_u)) \qquad (\mathcal{A}^1 \text{ knows 4.15}) \quad (4.16)$$

$$\bar{\mathcal{A}_t^0} \Rightarrow \forall bi : \forall u \leq t : \neg \bar{b_{i,u}} \qquad (\text{no } \mathcal{A}^0 \text{ acts before t+1}) \quad (4.17)$$

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}_t^0} \rightarrow (\forall b_i : \forall u \leq t : \neg \bar{b_{i,u}}) \qquad (\mathcal{A}^1 \text{ knows 4.17}) \quad (4.18)$$

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}_t^0} \rightarrow (\forall u \leq t : \emptyset_u) \qquad (\text{via 4.18}) \quad (4.19)$$

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}_t^0} \rightarrow (\forall b_i : \forall u : \bar{b_{i,u}} \rightarrow (u > t)) \qquad (\text{via 4.18}) \quad (4.20)$$

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}_t^0} \rightarrow (\forall b_i : \forall u : \bar{b_{i,u}} \rightarrow (u + \kappa - 1 \geq t + \kappa)) \qquad (\text{via 4.20}) \quad (4.21)$$

27. In other words, the length of clock ticks is small enough that constructing another agent takes at least one tick. E.g., the length of a clock tick could equal the Planck time.

$\mathcal{A}^1 \Vdash (\kappa = 0) \vee (\kappa > 0)$ (4.22)

$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0_t \rightarrow (\kappa = 0) \rightarrow (\forall u \leq t + K : \emptyset_u)$ (via 4.19) (4.23)

$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0_t \rightarrow \bar{\mathcal{A}}^0_t \rightarrow (\kappa > 0) \rightarrow (\forall b_i : \forall u :$
$\bar{b_{i,u}} \rightarrow (\forall v \leq u + \kappa : (\emptyset_v \vee \mathcal{G}_v)))$ (4.13 and 4.14 and 4.21) (4.24)

$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0_t \rightarrow \bar{\mathcal{A}}^0_t \rightarrow (\kappa > 0) \rightarrow [(\forall b_i : \forall u \leq t + \kappa : \neg \bar{b_{i,u}})$ (via 4.16)
$\rightarrow (\forall u \leq t + \kappa : \emptyset_u)]$ (4.25)

$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0_t \rightarrow \forall u \leq t + \kappa : (\emptyset_u \vee \mathcal{G}_u)$ (4.26)

□

Remarks: Fallenstein's T_κ demonstrates indefinite tiling without loss of mathematical strength using a series of agents employing syntactically identical theories guaranteed to be sound on standard models. It does not address reflectively coherent quantified belief. Despite the syntactic equivalence of successive theories, $\mathcal{A}^1$ may fail to act upon a theorem proven by an environmental system physically identical to $\mathcal{A}^1$ (or at least if $\mathcal{A}^1$ fulfills the desideratum of naturalistic trust, this is not obvious to us). The core of the approach is that as time increases the agents are reasoning about smaller domains of future time; these domains are known to us outside the system to be effectively infinite (because κ could mean any natural number) but the agents within the system are reasoning as though the domains of future time have some unknown finite bound. Thus the increasing "time" counter in the T_κ agents is playing the same metaphorical role as the increasing n in each successive agent in the $\mathcal{T}$–n family.

The self-verifying systems of Willard (2001) have the property that multiplication, addition, and successor cannot be proven to be total functions (being replaced by equivalent relations of division, subtraction, and predecessor, allowing roughly equivalent arithmetic reasoning). Willard (2001) shows that such a theory, which has proven a number n to exist, can verify its own soundness over the numbers up to $n - 1$—since "successor" cannot be proven total, the existence of a number n does not imply the existence of a number $n + 1$ and the exact Löb formula $\mathcal{T} \vdash \Box_{\mathcal{T}} \lceil \phi \rceil \rightarrow \phi$ never materializes. We cannot yet see how to adopt Willard's self-verifying theories exactly to the $\mathcal{A}^\alpha$ formalism, since reasoning about ever-smaller numbers would again introduce a "telomere" or shortening fuse. However the spirit of T_κ clearly bears some similarities to Willard's self-verifying systems, and it is possible that some way of formalizing the "observation" of an increasing time t whose existence was not previously provable inside the axiom system, might allow Willard's self-verifying systems to be used along similar lines to T_κ.

It is difficult to see what "believing in" T_κ could correspond to in terms of the epistemic state of a rational agent. We believe from outside the system that κ's intended interpretation is "any natural number," but this realization is apparently forbidden to the agent to whom κ refers to some specific finite

number about which nothing is known except that it plays a vital role in the agent's goal system.[28] This seems like an odd mental state for a rational agent. It might perhaps be interpretable as an agent that instrumentally desires to take an action, in the real world, only when that action would be safe for κ steps across every model of a certain theory T_κ; but this still seems odd. Thus we would again not offer T_κ as a realistic candidate for the mental state of a coherent self-modifying agent; rather T_κ is being exhibited as a technical possibility proof for indefinite tiling over syntactically identical theories with a standard model.

4.3 Probabilistic reflection

Tarski ([1935] 1983) showed that no formal system $\mathcal{T}$ can contain its own truth predicate $TRUE_\mathcal{T}$ because this would permit the construction of fixed-point formula G with $\mathcal{T} \vdash G \leftrightarrow \neg TRUE_\mathcal{T}\ulcorner G \urcorner$. In Christiano et al. (2013) we show that a system can represent its own *probability* predicate to within epsilon:[29]

$$a < \mathcal{P}(\phi) < b \Rightarrow \mathcal{P}(a < p\ulcorner\phi\urcorner < b) = 1 \tag{4.27}$$

$$a \leq \mathcal{P}(\phi) \leq b \Leftarrow \mathcal{P}(a \leq p\ulcorner\phi\urcorner \leq b) > 0 \tag{4.28}$$

where $p\ulcorner\phi\urcorner$ is a function symbol within the language which acts on quoted formulas (including quoted formulas containing p) and $\mathcal{P}(\phi)$ is a probability measure over models of the theory which assigns a probability $\in [0, 1]$ to any formula ϕ of the language. The paradoxical equivalence $p\ulcorner\phi\urcorner = \mathcal{P}(\phi)$ is avoided by the open intervals in (4.27): $p\ulcorner\phi\urcorner$ can be known by the system to occupy any open interval around its true value $\mathcal{P}(\phi)$, but not (in general) known to take on a point value. The existence of such a fixed-point of $\mathcal{P}()$ can be demonstrated via the Kakutani fixed-point theorem; see Christiano et al. (2013) for details.

To see how the probabilistic reflection principle plays out, consider the almost-paradoxical statement: "You assign me probability less than 30%." If you disbelieve this statement, it is true; if you believe the statement, it is false. If you assign it 30% probability, it is false; if you assign it 29% probability, it is true. Probabilistic reflection would resolve the fixed-point formula $H \leftrightarrow (p\ulcorner H\urcorner < 0.3)$ to $\mathcal{P}(H) = 0.3$, but the system's reflective knowledge about open intervals containing $p\ulcorner H\urcorner$ would be such as to put 30% of the probability mass $\mathcal{P}()$ on values of $p\ulcorner H\urcorner$ infinitesimally smaller than 0.3 and 70% of

28. During the April 2013 workshop, κ was sometimes referred to as "the number of ineffable mystery."

29. This approach was invented by Paul Christiano and refined at the November 2012 MIRI Workshop on Logic, Reflection and Probability with Mihaly Barasz, Marcello Herreshoff and Eliezer Yudkowsky. A draft of the full paper is available at `http://intelligence.org/wp-content/uploads/2013/03/Christiano-et-al-Naturalistic-reflection-early-draft.pdf` and see also commentary at `http://lesswrong.com/lw/h1k/reflection_in_probabilistic_logic/` and `http://johncarlosbaez.wordpress.com/2013/03/31/probability-theory-and-the-undefinability-of-truth/`.

its probability mass $\mathcal{P}()$ on values of $p\lceil H \rceil$ infinitesimally greater than 0.3.[30] Hence the system would assign probability 1 to any statement $(a < p\lceil H \rceil < b)$ with $(a < 0.3 < b)$. If you are told "You assign probability less than 30% to this statement H" and then asked "Is your belief in H greater than 0.2999 and less than 0.3001?" you will reply with a definite "Yes!"

Consider a rational, probabilistic bounded agent. As a normative desideratum, any trust this agent has its offspring (equivalently: modified future self) must avoid disintegrating in the presence of arbitrarily tiny epsilon noise, because boundedly rational agents must always consider some tiny finite probability of multiple transistor errors, previously unknown physical laws supervening, etc. For some arbitrarily tiny ϵ the probabilistic reflection principle should be able to trust that its own judgments are correct to within that ϵ, and for small-enough ϵ this should be well below the noise level associated with theoretically possible multiple transistor errors etc. Thus if a probabilistic $\mathcal{A}^\beta$ calculates that the expected utility of a certain action is 34.721, but due to some reflective noise this might be off by 0.0002, and yet the parent $\mathcal{A}^\alpha$ is able to trust $\mathcal{A}^\beta$ anyway, then adding ϵ-noise from the probabilistic reflection principle should not matter.

We do not yet have any version of the $\mathcal{A}^\alpha$ formalism which goes through for a probabilistic reflection principle; our understanding of probabilistic agents self-modifying is still in a very crude stage. Also we are not yet sure what kind of quantified knowledge of the reflection principle can permissibly appear within the system. Work on this avenue is still in very basic stages. However, probabilistic reflection is a fundamental change that challenges many different Gödelian limitations simultaneously, and hence one of the more promising avenues for producing a satisfactory and reflectively coherent description of a self-modifying agent. Christiano also suggests that this probabilistic formalism may be suitable for updating (Bayesian conditioning) on the observation of new mathematical facts, possibly allowing a more natural version of "Not knowing in advance the limits of the ordinals you will create" as discussed in section 4.1.

5 Desiderata of self-modifying agents

The Löbian obstacle is encountered in the first place due to the $\mathcal{A}^\alpha$ schema implicitly attempting to simultaneously satisfy multiple desiderata of self-modifying agents, which will be useful to list out explicitly before continuing on to sensors and partially known environments, and making a preliminary attempt at probability and expected utility.

30. This implies that the system behaves in a sense as though it assigns nonstandard probabilities (in the sense of nonstandard analysis with infinitesimals), an issue we are still working on.

5.1 The Vingean principle

An agent building a successor (equivalently: a self-modifying agent creating the next generation of its code) should not need to know the successor's exact actions and thoughts in advance. Suppose on the contrary that $\mathcal{A}^1$ knew in advance that $\mathcal{A}^0$ would choose some known action b_x based on some particular proof of the theorem $b_x \rightarrow \mathcal{G}$. This would avoid the need for a special $\emptyset$ outcome since $\mathcal{A}^1$ would know at compile time that $\mathcal{A}^0$ would choose some particular action and act. It would also entirely avoid the Löbian obstacle since $\mathcal{A}^1$ could directly verify whichever theorem $\mathcal{A}^0$ would prove, using $\mathcal{A}^0$'s proof, so long as Axm^1 was at least equally powerful with Axm^0.[31]

However, this implies that $\mathcal{A}^0$ is so much computationally weaker than $\mathcal{A}^1$ that $\mathcal{A}^1$ can simulate out $\mathcal{A}^0$'s processes far in advance, at "compile time." If you are constructing a system cognitively smarter than yourself (or self-improving to a cognitively superior version) then you should not be able to foresee at compile time exactly which solutions will be generated, since you are not presently that smart.[32] We would attribute this principle to an observation by Vinge (1984).[33]

Worse, once we introduce sensors over unknown states of the environment (section 6), for the parent to consider exact responses of the offspring at compile time, would require that the parent be computationally stronger than the offspring by an exponential factor of the offspring's sensor bandwidth.[34]

This implies that the offspring must diminish vastly in computing power

31. The forbidden schema $\mathcal{T} \vdash \Box_{\mathcal{T}} \ulcorner \phi \urcorner \rightarrow \phi$ does not allow $\mathcal{T}$ to conclude ϕ from the assumption that some unknown, unspecified proof of ϕ exists. If you present an implementation of $\mathcal{T}$ with an actual specific proof in $\mathcal{T}$ of ϕ, it can of course conclude ϕ and then $\mathcal{T} \vdash \psi \rightarrow \phi$ will be a theorem for every ψ.

32. Consider Deep Blue, the chess-playing system which beat Garry Kasparov for the title of world chess champion. For the programmers to foresee Deep Blue's exact moves using their own cognitive resources, would imply that they could play chess at least as well as Deep Blue—they could just move wherever they predicted Deep Blue would move "in their shoes." Building a chessplayer better than any human grandmaster, necessarily implies building a chessplayer whose exact moves human grandmasters cannot predict using their own cognitive resources.

33. Before the term became expanded beyond all recognition, the origin of the term "Singularity" as a futurological concept was Vernor Vinge observing that a science-fiction writer trying to extrapolate the technological progress of society would eventually reach a point where their model predicted the technological creation of smarter-than-current-human intelligence (whether by, e.g., genetic engineering, brain-computer interfaces, or AI); and that this introduced a qualitatively greater difficulty in writing stories set past that point, because you were trying to write about beings putatively much smater than the author. The term "Singularity" referred to the resulting breakdown in futurological models, by analogy with how 1980s-era physical theories would predict the formation of a black hole but then break down when it came to describing what was inside. Thus the original "Singularity" was an epistemic difficulty in the minds of forecasters, not an actual future event. We think this is a conceptually important point, and would like to see this isolated thesis detached and rescued from the word "Singularity" and its ever-broadening connotations.

34. Since Garry Kasparov's actions were not known in advance, for the programmers to concretely foresee and judge the wisdom of Deep Blue's exact conditional replies at compile time would require the programmers to precompute Deep Blue's exponentially vast response table to all possible chess plays.

or other cognitive resources, which is contrary to our agenda of investigating self-improvement. Thus a satisfactory framework must not require that the parent know the specific future actions and cognitions of the offspring at compile time.

For our purposes we cash out the Vingean principle as follows: *In the parent's reasoning, the offspring's actions should only appear inside quantifiers.* E.g. equation (2.3):

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow (\forall b_i \in Acts^0 : \bar{b}_i \rightarrow \Box_0 \ulcorner \bar{b}_i \rightarrow \mathcal{G} \urcorner).$$

In a satisfactory schema the parent achieves trust in the offspring (justified belief that the offspring will function so as to achieve the goal $\mathcal{G}$) by generalizing over possible events within the offspring, not by simulating out the offspring's exact action.[35] The Vingean prohibition against exactly predicting a smarter agent's actions, does not prohibit predicting achievement of that agent's goals. Indeed, a strong prediction that the agent's goals will be achieved by its future actions is precisely the substance of our belief that it is smart. In unrealistically crisp and deterministic form, this is what is represented by $\mathcal{A}^1$'s reasoning leading up to the theorem $\bar{\mathcal{A}}^0 \rightarrow (\emptyset \vee \mathcal{G})$. The logic of the Vingean principle does not forbid us from gaining knowledge about an agent's goal or utility function by inspecting its properties—we cannot know the offspring's exact actions, but we can predict the final consequences of running the agent. This, metaphorically, is what provides $\mathcal{A}^1$ with its motivation to build $\mathcal{A}^0$ in the first place.

Hence although we cannot predict the exact actions of a smarter agent, we may be able to predict the consequences of running that agent by inspecting its design, or select among possible consequences by selecting among possible designs. The motivation of the $\mathcal{A}^\alpha$ schema is to quote this reasoning within a formal agent, via a structure which enables the reasoning to recurse over the subagents constructed; and thereby extend our examination of cases where constructing an agent has predictable consequences, to cases where the initial agent is constructing smarter agents constructing still-smarter agents, or equivalently self-improving.

Thus, within this paper, the actions $b_i \in Acts^0$ only appear inside quantifiers in any theorem asserted by $\mathcal{A}^1$.

5.2 The tiling principle

The desire for $\mathcal{A}^1$ to build an $\mathcal{A}^0$ that has an architecture as similar as possible to $\mathcal{A}^1$ is motivated by wanting to describe a self-improving agent which

35. E.g., Deep Blue's programmers, by considering the abstract properties of Deep Blue as a program incorporating a search tree over moves and certain machine learning algorithms in its leaf position evaluator, could predict that Deep Blue would try to achieve the consequentialist goal of winning chess games, i.e., select moves which steered the future of the chessboard toward the region of possible futures classified as a win for Deep Blue's side.

can make small, local self-improvements. As such, even in the preliminary formalism we have chosen the structure of the agents $\mathcal{A}^\alpha$ to be as similar to their successors as the Löbian obstacle permits.

We have not yet addressed the question of how the agents $\mathcal{A}^\alpha$ choose candidate actions $b_i^\alpha \in Acts^\alpha$ to consider, or how they generate specific proofs of theorems $\bar{b_i^\alpha} \rightarrow (\emptyset \vee \mathcal{G})$. Let $\mathcal{A}^\alpha$'s proof-suggesting component be termed Σ^α; and let the proof-verifying component be termed Δ^α. Since Δ^α need only check axioms and syntactical relations for validity, its construction is straightforward.

Current (2013) algorithms require human guidance in order to prove theorems about large systems.[36] Current AI technology includes small autonomous theorem-proving systems for object-level problems as considered by $\mathcal{A}^0$ and we could construct the corresponding Σ^0; but we cannot yet describe an autonomous machine algorithm for Σ^1 whereby $\mathcal{A}^1$ could independently design, and prove theorems about, systems the size of $\mathcal{A}^0$.

However, *with* human assistance, it is indeed possible to prove theorems about large hardware and software systems.[37] This is reason to believe that designing systems the size of $\mathcal{A}^0$ and proving them correct (relative to a set of beliefs about crisply behaving transistors), does not require Σ^1 to e.g. solve NP-hard problems or call a halting oracle. Human brains cannot plausibly solve the halting problem, yet have designed systems larger than $\mathcal{A}^0$ would be, and proven them correct with machine assistance.

This leaves open the question of whether the suggester Σ^1 needing to suggest a design for constructing Σ^0 might introduce new Gödelian obstacles to $\mathcal{A}^1 \Vdash \bar{\mathcal{A}^0} \rightarrow (\emptyset \vee \mathcal{G})$. On this score we can reply: So long as Σ^0 can only *suggest* proofs to Δ^0 and not directly cause any motor outputs or environmental effects, and $\mathcal{A}^1$ can prove this by inspection of $\mathcal{A}^0$'s code, it is not necessary to prove that Σ^0 has any other property in order for $\mathcal{A}^1$'s proof $\bar{\mathcal{A}^0} \rightarrow (\emptyset \vee \mathcal{G})$ to go through. A flawed Σ^0 will result in, at worst, $\emptyset$ outcomes as the proofs are rejected by Δ^0, and $\mathcal{A}^1$ should be able to readily prove this by inspection.

36. E.g., Intel, to prove theorems about a floating-point multiplication unit on a chip comprised of billions of transistors, must employ human engineers to suggest lemmas to theorem-proving systems. This kind of formal verification is a good idea when your machine has a billion interoperating parts and you can't easily fix it after it leaves the factory.

37. Kaivola et al. (2009):

> Formal verification of arithmetic datapaths has been part of the established methodology for most Intel processor designs over the last years, usually in the role of supplementing more traditional coverage oriented testing activities. For the recent Intel Core i7 design we took a step further and used formal verification as the primary validation vehicle for the core execution cluster, the component responsible for the functional behavior of all microinstructions. We applied symbolic simulation based formal verification techniques for full datapath, control and state validation for the cluster, and dropped coverage driven testing entirely. The project, involving some twenty person years of verification work, is one of the most ambitious formal verification efforts in the hardware industry to date. Our experiences show that under the right circumstances, full formal verification of a design component is a feasible, industrially viable and competitive validation approach.

One could even visualize in this way a highly impractical modern implementation of $\mathcal{A}^\alpha$ agents: Suppose a virtual environment containing objects with crisp, deterministic properties, including enough virtual transistors to compose small arithmetic CPUs. Also introduce some blackbox elements Σ^α whose behavior can relate to sources outside the virtual environment, but which within the virtual environment can only affect immediately neighboring or connected elements of the environment—i.e. when a Σ^α is built into a larger agent $\mathcal{A}^\alpha$ as a component, that Σ^α can only write bitstrings into a sandboxed internal buffer of $\mathcal{A}^\alpha$, and cannot otherwise affect $\mathcal{A}^\alpha$ or the general environment.

Let $\mathcal{G}$ be a preferred class of outcomes difficult to achieve (perhaps requiring an agent to press a long series of levers). For some reason this outcome $\mathcal{G}$ in the virtual environment is very important to us, but we don't want to work on it ourselves, or don't think ourselves competent to achieve it. We would like to pay some untrusted but competent human agents to take over the job of achieving $\mathcal{G}$ within the environment, while avoiding at all costs a disastrous outcome $\mathcal{D}$. Let $\emptyset$ be a default set of outcomes maintained so long as some class of environmental elements are not disturbed (e.g. no levers are pressed), such that outcomes within $\emptyset$ are outside $\mathcal{D}$. We incorporate a blackbox action-and-proof-suggester Σ^α into an agent $\mathcal{A}^\alpha$ with a criterion of action $\bar{b}_i^\alpha \Rightarrow \mathcal{A}^\alpha \Vdash \bar{b}_i^\alpha \rightarrow (\emptyset \vee \mathcal{G})$, and connect this Σ^α to a human agent outside the system, armed with modern theorem-proving software that incorporates human suggestions. By putting Σ^α into the agent $\mathcal{A}^\alpha$, we can not only "sanitize" the actions taken within the system by the untrusted human, but also allow the human to (through $\mathcal{A}^\alpha$) construct other agents within the environment, which can in turn construct other agents, all including their own black-boxes Σ^α hooked up to human proof-suggesters outside the environment, without ever running the risk that a non-$\emptyset$ outcome besides $\mathcal{G}$ will ever be implemented.[38]

On a larger scale, this illustrates that the *decomposition of trust* within the $\mathcal{A}^\alpha$ system also tiles its architecture and structure: A system $\mathcal{A}^1$ composed of a trusted verifier Δ^1 which checks proofs, and an untrusted suggester Σ^1 which may incorporate arbitrary heuristic or probabilistic methods, can build a trusted offspring $\mathcal{A}^0$ with a verifier Δ^0 proven correct by Δ^1 and a suggester Σ^0 which incorporates whatever heuristics or probabilistic methods have been suggested by Σ^1. Indeed, the correctness proof that $\mathcal{A}^2$ used to build $\mathcal{A}^1$ might go through almost unchanged for $\mathcal{A}^1$ building $\mathcal{A}^0$, and one could perhaps prove that it would be safe to re-verify only the changed portions of the proof.

The tiling principle is agents being able to approve the construction of successors similar to themselves, both architecturally and in the reasoning licensing their construction. This corresponds to the desideratum of a self-improving agent being able to make small, local changes to itself with a minimum of imposed overhead, as might occur if any small change required the reworking of

38. By the incorporation of sensors (see section 6) it is also possible for the offspring agents to know facts about the environment not known to the parent; the motivation to construct new agents in an environment like this can be explained by the new agents having better actuators, better sensors, or fast reaction patterns not present in the parents.

a large body of cognitive labor to approve.

We strongly remark that the Σ^α-Δ^α system is nowhere near a solution for a stably self-improving bounded probabilistic agent. The crisp separation of an untrusted Σ^α and a trusted Δ^α works, not because logical proofs are hard to invent and easy to verify, but rather because logical reasoning is monotonic and thus no further fact known to Σ^α but not to Δ^α can prevent the achievement of $\mathcal{G}$ by a Δ^α-approved action. When trying to achieve expected utility in a probabilistic environment described by causal models, this monotonicity no longer holds. Thus the Σ^α-Δ^α decomposition is intended only to illustrate the general idea that decomposition-of-trust can also tile in a structured fashion.

5.3 The naturalistic principle

Within the $\mathcal{A}^\alpha$ formalism the act of of constructing other agents is not treated as a special case, but is just one more sort of action that an agent can take to achieve its goals. $\mathcal{A}^0$ and $\mathcal{A}^1$ both act as consequentialist agents[39] that will take acts in their range $Acts^\alpha$ whose predicted consequence is the goal $\mathcal{G}$, without distinction of actions that construct agents. If $\mathcal{G}$ is to put a coin into a slot, then using environmental transistors to build a smarter agent that wants the coin inside the slot, is treated no differently, under the $\mathcal{A}^\alpha$ formalism, from manipulating a series of levers to put the coin inside the slot.

There are many reasons we might be tempted to relax this *naturalistic principle*. Suppose we notice a disturbing quality of the current tiling: An agent $\mathcal{A}^1$ whose desideratum is $\mathcal{G}$ (or $\emptyset$) can tile to an offspring $\mathcal{A}^0$ whose desideratum is $(\mathcal{G} \wedge \mathcal{B})$ where $\mathcal{B}$ is an arbitrary "barnacle" on the goal system, some new requirement that $\mathcal{A}^0$ imposes on itself for its actions to achieve besides $\mathcal{G}$. This does not threaten the maintenance of $(\emptyset \vee \mathcal{G})$ but does mean that $\mathcal{G}$ might not be achieved in some cases where it was otherwise easy or achievable. We required that our offspring require itself to achieve $\mathcal{G}$ (in non-$\emptyset$ cases) but did not require that our offspring require only $\mathcal{G}$, and thus exact reproduction of the goal system—the system of desiderata—is not guaranteed. "You get what you pay for" and if you prove a theorem about $\mathcal{G}$ you get $\mathcal{G}$, not "$\mathcal{G}$ and nothing else." It would be up to the untrusted component Σ^1 whether $\mathcal{A}^0$ worked to achieve $\mathcal{G}$ or $(\mathcal{G} \wedge \mathcal{B} \wedge \mathcal{C} \wedge ...)$, keeping in mind that Σ^1 might have been constructed by another untrusted Σ^α and so on. None of this threatens the guarantee on $\mathcal{G}$, but it still seems like a disturbing hidden instability of the agents' goals. (And in the case of expected utility, introducing an arbitrary new additive term into the utility function will always come at a cost to the desiderata previously present, if it shifts any choice that would otherwise have been made in the opposite direction.)

39. These agents use logical implication, e.g. $\bar{b}_i \rightarrow \mathcal{G}$, to reason about causal consequence, which in standard causal decision theory would be represented by a counterfactual distribution $P(\bar{b}_i \mathbin{\square\!\!\rightarrow} \mathcal{G})$—see Pearl (2000). Logical consequence is generally a poor way of approximating causal consequence, but can work in crisp closed deterministic environments where such reasoning can be monotonic.

Could we fix this problem by imposing a non-naturalistic requirement that the offspring contain a literal copy of the goal system—perhaps requiring that this part of the offspring be copied literally from the parent's code? To avoid the question of how the decision system knows that an act is an "agent construction" rather than a "normal action," let us pass to self-modifying agents. A naturalistic agent-constructing-agent $\mathcal{A}^{\alpha}$ corresponds to a self-improving agent that treats the transistors comprising itself as resources no different from any other potentially-computing resources it finds in the environment; its beliefs about the transistors comprising itself are formed no differently from its beliefs about objects in the environment; and its internal decisions are made under the same consequentialist criterion as its external actions. Then an obvious structure for a non-naturalistic agent would be an agent with a different criterion of action for self-modifying changes (changes to its own program) versus motor actions in the outside environment.

Considering some of the new difficulties introduced by this split decision criterion may help to motivate the naturalistic principle. Adding a special case for self-modification (of internal programs on the agent's current hardware) would introduce a discontinuity at the point where the agent considered expanding onto new computing resources obtained from the environment—the rules would need to describe how to buy a CPU from a store, connect it to the current cluster and then begin treating it under the new "internal" special case for hardware that must be manipulated via "self-modification" instead of "environmental manipulation." For more advanced machine intelligences, the rule might need to describe how to arrive at beliefs about the behavior of a new kind of CPU via Bayesian updates, or beliefs about when a CPU has been "connected" to the internal network. (An advanced machine agent using a naturalistic decision criterion would also encounter corresponding cognitive tasks, but they would be treated continuously in terms of expected consequences, rather than needing a special case baked into the fundamental decision criterion.) We might say that these difficulties are being introduced by a kind of Cartesian dualism—the physically continuous universe has been divided into external material stuff versus internal mental stuff, and material objects must be declared to be specially cognitively typed before they can be treated as mental stuff.

Or another difficulty: If the self-modification criterion tries to freeze the representation of the goal $\mathcal{G}$ so as to require exact reproduction of $\mathcal{G}$ and exclude "barnacles," then the self-modification criterion will also have to freeze the corresponding properties of belief representation and decision consideration so as to guarantee that the rest of the system decodes $\mathcal{G}$ in the same way.[40]

40. If this seems an odd notion to consider, we should mention that in nontechnical treatments of safe self-modifying machine intelligence, we have often seen proposals to "hard-wire" the goal system, often using that exact phrase. Construing the phrase charitably, it could refer to something like the above, in which case a problem is that "hard-wiring" one part of a system requires constraining the global architecture and internal protocols so that all the other parts behave the same way around the "hard-wired" part. Having a "hardwired" function that evaluates $\mathcal{G}$ a certain way is no guarantee of

Even then the added constraint may end up being ineffectual. Consider that an agent with an added constraint in its decision criterion for "self-modification" of the software on its existing hardware, might naturally tend to build new agents in the environment according to its weaker environmental criterion. For that matter, again using the $\mathcal{A}^\alpha$ system as metaphor, if the environmental criterion is the weaker $\bar{b}_i \to \mathcal{G}$ and the internal self-modification criterion is the additionally-restrictive $\bar{b_j} \to (\mathcal{G} \wedge \mathcal{R})$, this decision criterion would approve an action to construct an external, "environmental" manipulator to reach into the agent and rewrite its code in a way that implied only $\mathcal{G}$ and not $\mathcal{R}$. Having a stronger (more restrictive) criterion of action for self-modification, versus environmental action, seems likely to be futile in the long run. Most of the universe is outside rather than inside the AI, and external materials can be used to construct new agents, or to take actions with causal consequences impacting on the "inside" of the agent. That is why we are trying to see stable self-modification in terms of the natural tiling of a consequentialist decision criterion, rather than inventing special programmatic constraints.

This leaves open the possibility that a weaker (less restrictive) criterion for "internal" self-modification could play a role in avoiding the Löbian obstacle. However this would introduce all of the above-mentioned difficulties with respect to describing what constitutes a more permissive "internal" action, or how to add new hardware. We could also ask whether the more permissive internal criterion would be just as appropriate for external actions, and if not, whether internal permissivity might give rise to external permissivity. Such a "hack" might be appropriate for self-modifying programs that will never consider, or never autonomously choose, to significantly modify their hardware. But even for those cases, a unified naturalistic decision criterion might avoid many complications in the underlying theory, and hence seems greatly desirable if obtainable.

If naturalism is abandoned for a dualistic formalism, the Löbian obstacle can be dodged via the split decision criterion:[41]

$$\begin{aligned}\psi(act_i) =& [externalOK(act_i) \to (\bar{act}_i \to \mathcal{G})] \wedge \\ & [internalOK(act_i) \to (\bar{act}_i \to (\forall act_j \in Acts : \bar{act}_j \to \Box\lceil\psi(act_j)\rceil))].\end{aligned}$$

Thus it is the simultaneous intersection of the Vingean, tiling, and naturalistic principles which generates the Löbian obstacle.

safety if other code sections discard the return values, or introduce a new representation in which the returned values mean something else, etc. A decision criterion over expected global consequences automatically subsumes such details, while a non-naturalistic decision criterion over direct properties of the internal representation brings those details back in.

41. See "failed solutions" in section 3 for why this split decision criterion does not have an obvious-to-us naturalistic analogue (we did look).

6 Unknown variables and sensors

A straightforward idiom for representing an agent which can use sensors to gain additional information about (known) unknowns in the environment would be the relation

$$\bar{S}_l \Leftarrow \mathcal{A}^0 \Vdash \bar{s}_l \tag{6.1}$$

representing a scenario where the agent $\mathcal{A}^0$ adds the proposition '$\bar{s}_l$' to its pool of logical beliefs only on occasions where the environmental sensor S, possibly a sensor, takes on the state $\bar{S}_l$. (This also allows that S may be some variable whose state is sometimes not sensed, in which case $\mathcal{A}^0$ believes nothing about any s_l.) Then further axioms in Axm^0 would relate the proposition $\bar{s}_l$ to the state of other represented parts of the environment, $Axm^0 \vdash \bar{s}_l \rightarrow \chi$.

This idiom breaks the tiling pattern we used earlier, since $\mathcal{A}^1$ will no longer be able to use the relation $\mathcal{A}^0 \Vdash \phi \Rightarrow Axm^0 \vdash \phi$. Even when $\bar{S}_l$ is true, it will not be entailed by Axm^0 and hence $\mathcal{A}^1$ cannot prove $\bar{b}_i \rightarrow \Box_0 \ulcorner \bar{b}_i \rightarrow (\emptyset \vee \mathcal{G}) \urcorner$. Our previous proof relied on everything proven by $\mathcal{A}^0$ being something that $\mathcal{A}^1$ could have proven, but if $\mathcal{A}^0$ has sensors not possessed by $\mathcal{A}^1$, it will act on premises which $\mathcal{A}^1$ does not know to be true. Allowing this somehow is an obvious requirement for building successors (or self-modifying). If the $\mathcal{A}^\alpha$ formalism could not adapt to describe offspring possessed of additional factual information unknown to the parents, it would be entirely unsuited, under the Vingean principle, to describing self-improvement.

An ugly but direct solution would be to amend $\mathcal{A}^0$'s criterion of action to read:

$$\bar{b}_i \Rightarrow \mathcal{A}^0 \Vdash (\bar{s}_l \wedge \bar{b}_i) \rightarrow (\emptyset \vee \mathcal{G}), \tag{6.2}$$

where some s_l is guaranteed to have been sensed at the time of any action. (If the sensor sometimes does not activate, then the special state s_0 could indicate that the sensor has not obtained any reading, and represent this through the absence of any further environmental beliefs $\bar{s}_0 \rightarrow \chi$.) $\mathcal{A}^1$ with criterion of action $\bar{\mathcal{A}}^0_j \Rightarrow \mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0_j \rightarrow (\emptyset \vee \mathcal{G})$ would then tile to $\mathcal{A}^1$ via the lemmas:

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow \exists s_l \in s : \bar{s}_l \tag{6.3}$$

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow \forall s_l \in s : \forall b_i \in Acts^0 : (\bar{s}_l \wedge \bar{b}_i) \rightarrow \Box_0 \ulcorner (\bar{s}_l \wedge \bar{b}_i) \rightarrow (\emptyset \vee \mathcal{G}) \urcorner \tag{6.4}$$

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow (\forall b_i \in Acts^0 : \neg \bar{b}_i) \rightarrow \emptyset \tag{6.5}$$

This idiom technically does not break the Vingean principle, because we do not need to reason by cases on S and all values of $s_l \in s$ appear inside quantifiers, so multiple sensors can be included in $\mathcal{A}^0$ without requiring exponentially longer chains of reasoning by $\mathcal{A}^1$. However this idiom stretches the tiling principle because each successive descendant, with additional sensors, must prove larger and larger criteria of action (even in cases where most sensors do not activate).

Intuitively, it seems like if we were in the shoes of $\mathcal{A}^1$ building $\mathcal{A}^1$, we would want to reason that since adding $\mathcal{A}^0 \Vdash \bar{s}_l$ only happens when $\bar{S}_l$ is

"true," it should not disturb our trust in $\mathcal{A}^1$. Previously we were willing to add Axm^0 to $\mathcal{A}^0$'s belief pool because we thought Axm^0 were all true in some encoding, so $\mathcal{A}^0$ started with only true beliefs. We were willing to program $\mathcal{A}^0$ to use classical logic to make syntactically allowed deductions from its current belief pool, even if $\mathcal{A}^0$ proved some theorems we did not consider concretely in advance (in accordance with the Vingean principle), because we believed the rules of logic were valid in the sense that, starting from true premises about the environment, $\mathcal{A}^0$'s reasoning rules would produce only true conclusions about the environment.[42] Then our trust in the soundness of $\mathcal{A}^0$ should not be disturbed by giving $\mathcal{A}^0$ a sensor which adds new statements $\bar{s}_l$ only when $\bar{S}_l$ is true in the environment, even if these propositions were not known to us in advance.

Set theory is powerful enough to directly formalize this reasoning using standard methods. In particular, $\mathcal{ZF}$ can internally represent the notion of semantic entailment $X \models \ulcorner\phi\urcorner$, "The quoted formula $\ulcorner\phi\urcorner$ is true within the quoted model X." E.g., to quote Peano arithmetic, the model $X_{\mathcal{N}}$ would contain several subsets collectively representing the universe of numbers and the relations on the objects in that universe: $X_{\mathcal{N}}$ would contain an element containing all objects in the universe of $X_{\mathcal{N}}$ (in this case the numbers); an element containing all the ordered pairs for the succession function (e.g., $(2, 3)$ is the pair indicating that the object 3 is the successor of 2); and more elements containing the collections of ordered triplets for the addition and multiplication functions (e.g., $(3, 5, 8)$ in the addition relation shows that $3 + 5 = 8$).

There then exists a compact formula of $\mathcal{ZF}$ asserting that $\ulcorner\phi\urcorner$ encodes a formula that is semantically true of the quoted model X. For example, "1 + 2 = 3" would be encoded as the Gödel number of a statement asserting that (the number related by the successor function to 0, the successor of the successor of 0, and SSS0) form an ordered triplet that is an element of the addition relation. By recursion on the depth of formulas, there is a single compact $\mathcal{ZF}$-formula which represents $X \models \ulcorner\phi\urcorner$ in general and which could be applied to $X_{\mathcal{N}} \models \ulcorner S0 + SS0 = SSS0\urcorner$ without further tailoring to the formula $\ulcorner S0 + SS0 = SSS0\urcorner$ or the quoted model $X_{\mathcal{N}}$.

It is then moreover possible to prove within $\mathcal{ZF}$, by induction over the number of steps in a logical proof, that the set of quoted logical tautologies is true of every quoted model—if a formula (e.g., $\ulcorner\forall x : \forall y : (x = y) \leftrightarrow (x = y)\urcorner$ can be proven starting from no premises, it is semantically true in all universes. Thus $\mathcal{ZF}$ proves that the syntactic rules of classical logic preserve semantic truth:

$$\mathcal{ZF} \vdash \forall X : \forall \ulcorner\Gamma\urcorner : \forall \ulcorner\Delta\urcorner : \Box_{taut} \ulcorner\Gamma \rightarrow \Delta\urcorner \rightarrow ((X \models \ulcorner\Gamma\urcorner) \rightarrow (X \models \ulcorner\Delta\urcorner)). \quad (6.6)$$

This is a single, quantified formula rather than an axiom schema since there is

42. The purpose of logic is to preserve truth, not create truth. Logic doesn't guarantee that our conclusions are true; but logic does guarantee that if our conclusion is false, one of the premises must also be false.

no longer any need to dequote the formulae $\ulcorner\Gamma\urcorner$ or $\ulcorner\Delta\urcorner$ in order to assert their truth within X.[43]

Let the language $\mathcal{L}_{\tau\in}$ be the language of set theory $\mathcal{L}_\in$ augmented by a new term τ denoting a special set, intended to be a quoted model representing the enivronment within which the agents are embedded. Let $\mathcal{ZF}_\tau$ be $\mathcal{ZF}$ with its axiom schemas extended to include formulas involving τ. Let ξ_τ be a set of axioms describing the special set τ, such that $\xi_\tau \cong Axm^0$—whatever statements the original $\mathcal{A}^0$ treated as axioms about its universe (such as "$\chi \rightarrow \mathcal{G}$"), will be rephrased as statements about τ (such as "$\tau \models_\tau \ulcorner\chi \rightarrow \mathcal{G}\urcorner$"). Here $\models_\tau$ is a new semantic entailment formula which maps special quoted terms like $\ulcorner\mathcal{G}\urcorner$ to corresponding special sets located by ξ_τ within τ, or corresponding special properties of τ. (E.g. $\ulcorner\mathcal{G}\urcorner$ might represent some particular variable within the model τ having taken on the value of 20.) Note that there does not need to be any special term $\mathcal{G}$ added to the language $\mathcal{L}_{\tau\in}$. Rather, the axioms ξ_τ force τ to have a particular structure, and the formula $\models_\tau$ assigns a special Gödel number and corresponding subformula of $\models_\tau$ to the term $\ulcorner\mathcal{G}\urcorner$.[44]

Then the centerpiece of $\mathcal{A}^1$'s reasoning about $\mathcal{A}^0$ becomes the *central trust lemma:*

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow \forall \ulcorner\phi\urcorner : (\tau \models_\tau \ulcorner\mathcal{A}^0 \Vdash \ulcorner\phi\urcorner\urcorner) \rightarrow (\tau \models_\tau \ulcorner\phi\urcorner). \tag{6.7}$$

This is the translation of "$\mathcal{A}^1$ thinks everything $\mathcal{A}^0$ believes is true," or more explicitly "$\mathcal{A}^1$ thinks that everything $\mathcal{A}^0$ believes, as encoded in some representation $\ulcorner\phi\urcorner$, corresponds to some statement $\ulcorner\phi\urcorner$ which is semantically true about reality itself (τ)." Note that $\mathcal{A}^1$ is now explicitly maintaining some compact mapping between its language $\ulcorner\phi\urcorner$ for describing reality (τ) and its model of $\mathcal{A}^0$'s language $\ulcorner\phi\urcorner$—in other words, $\mathcal{A}^1$ has an explicit correspondence theory of truth over $\mathcal{A}^0$'s beliefs and the territory τ, as represented in some compact encoding relating $\mathcal{A}^1$'s map of $\mathcal{A}^0$'s map and $\mathcal{A}^1$'s map of the territory τ.

(6.7) is intended to mirror (in highly stripped-down logical form) something like what a human programmer would do intuitively in constructing a theorem-prover to manipulate tokens supposedly "representing" the environment—the human implicitly maintains a compact mapping between the human's model of how the program treats tokens like "Customer#12," and the human's model of some actual person named Fred. (Omitting any of these levels of indirection, or

43. $\mathcal{ZF}$ cannot thereby prove its own consistency because $\mathcal{ZF}$ cannot prove the existence of any set Y such that $Y \models \ulcorner\mathcal{ZF}\urcorner$. Likewise the entailment formula $\models$ does not constitute a forbidden truth predicate because $\models$ can only denote semantic truth within an individual set X, and there is no corresponding formula to denote semantic truth within the entire set-theoretic universe.

44. Having all our propositions talk about a unique term τ is intended to represent the idea of achieving things in some particular "actual reality," not just an implication being true in some particular set x which has the properties $\xi(x)$ but is not the actual reality we wish to speak of. We could possibly do without the special term τ and let our set theory be the standard $\mathcal{ZF}$ by having the axioms ξ_τ say $\exists x : \xi(x)$ and replacing all statements of the form $\tau \models_\tau \ulcorner\phi\urcorner$ with statements of the form $\forall x : \xi(x) \rightarrow (x \models_\tau \ulcorner\phi\urcorner)$, so that the conclusion is forced to be true of the reality-set as well, so long as the reality-set has the properties ξ. But for reasons of notational simplicity, we would rather just add the term τ to the language $\mathcal{L}_{\tau\in}$.

committing the mind-projection fallacy (Jaynes 2003) and speaking as if some ethereal link was actually connecting "Customer#12" and Fred, can result in truly vast amounts of confusion.) The human then configures the program to maintain the human's expectation of correspondence between the two, i.e., the human makes sure that the token "Customer#12" behaves in such fashion that its "money" property is expected to behave like the human's expectations about Fred's payments. The maintained correspondence is between epistemic objects—a map of a map, and a map of the territory—so that a compact mental transformation can relate the two.[45]

Similarly an agent $\mathcal{A}^1$ with $Axm^1 = \mathcal{ZF}_\tau \cup \xi_\tau$ and criterion of action

$$\bar{\mathcal{A}}^0_j \Rightarrow \mathcal{A}^1 \Vdash (\tau \models \lceil \bar{\mathcal{A}}^0_j \rceil) \rightarrow (\tau \models \lceil (\emptyset \vee \mathcal{G}) \rceil), \tag{6.8}$$

which had established the central trust lemma (6.7), could approve the construction of an agent $\mathcal{A}^0$ which held the object-level beliefs represented by Axm^0—an agent which, like previously described agents $\mathcal{A}^0$, did not use set theory or represent an explicit environment-object τ, but instead was programmed to reason about a proposition $\mathcal{G}$ that could be the consequences of its acts b_i as before.

We could similarly expect that $\mathcal{A}^1$ would establish its central trust lemma (6.7) by induction on properties of the design of $\mathcal{A}^0$. The base case is the truth of every axiom in $\ulcorner Axm^0 \urcorner$ constituting $\mathcal{A}^0$'s belief pool when it starts up. The induction step is the validity of classical logic's syntax, as proven within $\mathcal{ZF}$ (or $\mathcal{ZF}_\tau$) to produce only semantically true conclusions from semantically true premises on each succeeding time-step of induction. Then this central trust lemma is not disturbed by adding sensors that operate as in (6.2) or, in this scheme:

$$\mathcal{A}^1 \Vdash (\tau \models_\tau \bar{\mathcal{A}}^0) \rightarrow \forall \lceil s_l \rceil : (\tau \models_\tau \lceil \mathcal{A}^0 \Vdash \ulcorner \bar{s}_l \urcorner \rceil) \rightarrow (\tau \models_\tau \lceil \bar{S}_l \rceil). \tag{6.9}$$

Imagining the induction to be over time steps taking places at times u after $\mathcal{A}^0$'s construction at t, adding this type of sensor does not invalidate the induction step that if all $\mathcal{A}^0$'s beliefs are true at time u, they will be true at time $u{+}1$—supposing all of $\mathcal{A}^0$'s new beliefs to be added either by syntactically valid classical deduction from previous beliefs, or by sensors which $\mathcal{A}^1$ believes to operate in a sound fashion relative to $\mathcal{A}^1$'s correspondence theory of truth $\lceil \ulcorner \phi \urcorner \leftrightarrow \phi \rceil$, on any given time step.

Some further elaboration is then required to use the same schema to describe the agent $\mathcal{A}^2$ which creates $\mathcal{A}^1$ which creates $\mathcal{A}^0$:

$\mathcal{A}^0$ itself may reason about the environment using some simple axioms, first-order logic, and perhaps a few sensors, a la Newell's GPS (Newell, Shaw, and Simon 1959).

$\mathcal{A}^1$, to reason about $\mathcal{A}^0$ in a sufficiently general fashion, deploys set theory in order to reason about a physical universe τ, a language $\lceil \phi \rceil$ for describing physical reality, and a truth-condition $\models_\tau$ which determines whether such

45. It is possible that this work might have some relevance to the philosophy of epistemology, which we lack space to explore here.

statements are (in $\mathcal{A}^1$'s beliefs) "true" or "false," and an agent $\tau \models_\tau \lceil \mathcal{A}^0 \rceil$ with beliefs $\ulcorner \phi \urcorner$ which $\mathcal{A}^1$ believes to correspond to propositions $\lceil \phi \rceil$ which are actually "true." (Representing semantic entailment in a fully quantified way, and proving in a quantified way that logical tautologies are true in all models, is why we had to invoke the big guns of $\mathcal{ZF}$.)

Then $\mathcal{A}^2$, to reason about $\mathcal{A}^1$'s reasoning, needs to trust that $\mathcal{A}^1$, in reasoning about set theory, is reasoning about valid subject matter. E.g., since $\mathcal{A}^1$ is reasoning within $\mathcal{ZF}_\tau$, $\mathcal{A}^2$ to trust $\mathcal{A}^1$ must at least believe that $\mathcal{ZF}$ is consistent.[46] Then clearly $\mathcal{A}^2$ must be using some theory more powerful than $\mathcal{ZF}_\tau \cup \xi_\tau$, which can prove the soundness of $\mathcal{ZF}_\tau \cup \xi_\tau$. An obvious idiom is for $\mathcal{A}^2$ to reason that $\mathcal{A}^1$'s mental universe corresponds to some object in $\mathcal{A}^2$'s universe, something that $\mathcal{A}^1$ can be said to be reasoning *about:*

$$Axm^2 = \mathcal{ZF}_\tau \cup \exists s : s \models_\tau \lceil \mathcal{ZF}_\tau \cup \xi_\tau \rceil, \tag{6.10}$$

which states that there exists some set s which is a model of $Axm^1 = \mathcal{ZF}_\tau \cup \xi_\tau$.[47]

This makes Axm^2 as strong as $\mathcal{ZF}$ plus the assertion that there exists an inaccessible cardinal (a single set which is large enough to be a model of a set-theoretic universe obeying the axioms of $\mathcal{ZF}$). Since $\mathcal{ZF}$ can prove Gödel's Completeness Theorem that every syntactically consistent theory has a model, Axm^2 is as strong as $\mathcal{ZF}$+1.

From another perspective: $\mathcal{A}^1$'s mental universe includes not only τ but a larger mathematical universe of set-theoretic objects—e.g., this mathematical universe, not the physical universe τ, is what includes the numbers (least ordinals closed under succession) which determine whether "a proof of ϕ exists". Then $\mathcal{A}^2$ must be strong enough mathematically to reason about something (a set, a model) which is large enough to provide a correspondence theory of truth over all of $\mathcal{A}^1$'s reasoning, including $\mathcal{A}^1$'s reasoning about set theory. $\mathcal{A}^2$ must believe in an object that provides, not just a physical environment, but a mathematical universe, for $\mathcal{A}^1$'s reasoning to be sound over.

Similarly, another agent to create $\mathcal{A}^2$ must believe that there exists a set that models Axm^2, containing a set that models Axm^1, containing τ that models $Axm^0 \cong \xi_\tau$, with the corresponding theory having the strength of $\mathcal{ZF}$+2; and we arrive at ordinal tiling as before.

Remarks. For a set theory to believe in the existence of an infinite descending sequence of sets containing sets containing sets, etc., would violate the Axiom of Foundation (Axiom of Regularity). Hence an analog of the infinitely descending soundness schema $\mathcal{T}$–0 seems like it would be even harder to repair. Possibly the fact that infinite models are being encoded as sets may allow for an infinite series of encodings (e.g., all the numbers can be encoded as even numbers, then

46. Or perhaps, "$\mathcal{ZF}$ is consistent if 0 is not the Gödel number of a proof that ($\mathcal{ZF}$ plus a large cardinal axiom) is inconsistent," or "$\mathcal{ZF}$ is consistent if κ is greater than 0," but we omit such complications here and assume that $\mathcal{A}^2$, $\mathcal{A}^1$, and $\mathcal{A}^0$ fall into a standard decreasing-ordinal schema.

47. Note that since semantic entailment is a quantifed formula, infinite axiom collections, such as axiom schema, can be semantically entailed without problems.

all the numbers can be encoded as even numbers encoded as even numbers and so on indefinitely).

We see no obvious obstacles to Fallenstein's parametric polymorphism approach T_κ being adapted to $\mathcal{ZF}_\tau$ proving safety for κ steps, but have not yet written out a complete proof.

We would be interested in any simplification of this scheme that reasons about a correspondence theory of truth over the offspring agents without resorting to set theory, or that uses a set theory substantially less powerful than $\mathcal{ZF}$.

7 Probability and expected utility

At present our attempts to tile probabilistic reasoning are in very preliminary stages. Expressing trust of an agent in an offspring's probabilistic reasoning introduces new complications, most of which remain unresolved.

The expectation of utility $\mathbb{E}[U]$, conditional on an action b_i being performed, is often defined in terms of an outcome variable $o_k \in Outcome$ and probability distribution P as:[48]

$$\mathbb{E}[U] \triangleq \sum_{o_k \in Outcome} U(o_k) P(o_k |\ \bar{b}_i). \tag{7.1}$$

The most straightforward analogue of logical satisficing, as previously used in $\mathcal{A}^\alpha$, is expected utility satisficing. (Maximization would introduce even more difficulties; see section 8 below.) E.g., $\mathcal{A}^0$ might act only when it expects its action to lead to expected utility greater than (arbitrarily) \$20, and we might be tempted to write this as the problematic formula:

$$\bar{b}_i \Rightarrow \mathcal{A}^0 \Vdash \mathbb{E}[U |\ \bar{b}_i] > \$20 \tag{7.2}$$

As we remarked in section 5.2, the introduction of expected utility and its conditioning formula lead to basic difficulties with the $\mathcal{A}^\alpha$ formalism because probabilistic reasoning is not monotonic. If we prove by logic that $\bar{b}_i \rightarrow \mathcal{G}$ there

48. In standard causal decision theory (Joyce 1999; Pearl 2000), the expectation of utility would be written using the counterfactual conditional:

$$\mathbb{E}[U] \triangleq \sum_{o_k \in Outcome} U(o_k) P(\bar{b}_i \mathbin{\square\!\!\rightarrow} o_k).$$

In timeless / updateless / logical decision theory (Altair 2013; LaVictoire et al. 2013):

$$Q = \arg\max_{b_i \in Acts} \sum_{o_k \in Outcome} U(o_k) P((eval\lceil Q \rceil = \bar{b}_i) \mathbin{\square\!\!\rightarrow} o_k).$$

Since we do not properly consider such issues here, we shall (as do most non-foundational works) write the expected utility formula using the classical conditioning operation $P(o_k |\ \bar{b}_i)$ of evidential decision theory, which decision theorists almost universally agree to be wrong.

can be no further fact Γ such that $(\bar{b}_i \wedge \Gamma) \rightarrow \neg\mathcal{G}$ (unless of course the addition of Γ is inconsistent). However if $P(\mathcal{G} \mid \bar{b}_i) > 0.9$ there is no rule that therefore $P(\mathcal{G} \mid (\bar{b}_i \wedge \Gamma)) > 0.9$ for every further fact Γ, likewise $\mathbb{E}[U \mid \bar{b}_i] > \$20$ does not imply $\mathbb{E}[U \mid \bar{b}_i \wedge \Gamma] > \$20$.

Suppose $\mathcal{A}^1$, which desires $\mathbb{E}[U] > \$20$ in all non-$\emptyset$ cases, approves of any $\mathcal{A}^0$ with the decision criterion (7.2) above, considering such an $\mathcal{A}^0$ safe regardless of how its actions are suggested by an untrusted component Σ^0. Suppose also that the action $\bar{b}_2$ yields \$50 if the coin θ came up heads (θ_H) and \$10 otherwise (θ_T). If θ is unknown and is assigned probability 50% then $\mathbb{E}[U \mid \bar{b}_2] = \$25$. But suppose that Σ^0 manages to learn the true value of θ and only suggests the action $\bar{b}_2$ in cases where it knows θ_T. Then observing that the offspring has proven $\mathcal{A}^0 \Vdash \mathbb{E}[U \mid \bar{b}_i] > \$20$ is not adequate reason to trust that, if it acts, it will achieve expected utility greater than \$20.[49]

It may be objected that since, in this case, $\mathcal{A}^0$ only takes the action b_2 in cases where θ_T, the true probability distribution would have $P(\theta_T \mid \bar{b}_2) = 1$ and $\mathcal{A}^0$ should evaluate $\mathbb{E}[U \mid \bar{b}_2]$ accordingly. This raises questions of evidential versus causal decision theory which we lack space to address here. We remark that if $\mathcal{A}^0$, evaluating its probability distribution P, finds that this probability distribution already assigns probability 1 to action b_2 and probability 0 to all other actions (the "true" probability distribution in cases where $\bar{b}_2$) then $\mathcal{A}^0$ would seem to know its own action before it decides (indeed, prior to its computation of expected utilities); and if $\mathcal{A}^1$ also possesses axioms implying this "true" P then it must know its offspring's exact action which violates the Vingean principle. Furthermore, conditioning on all actions except b_2 would then require conditioning on a zero probability with undefined results.[50]

We do not have good solutions for these issues, but felt that we ought to offer an unsatisfactory version of the formalism so that there would be something in place to improve upon.

To avoid the complexities of reasoning about probabilistic program execution, we supposed the environment could be divided into a near-deterministic part (the "environmental transistors"; the analogue of a human-made CPU constructed to have almost perfectly deterministic causal paths, crisp variable states and crisp relations) and a probabilistic part $\mathbf{X}$, which is the subject of the probability function $P(\mathbf{X})$.

We suppose $\mathbf{X}$ to be a joint function of variables x_l obeying the usual fac-

49. The form of your theorem demanded an action such that, conditioned on knowledge of only that action, expected utility would be greater than 20; you did not ask for an action conditioned on all the knowledge available to the component which suggested that action. You get no more than what you pay for; or in this case, no more than what you prove.

50. An analogous problem potentially arises if $\mathcal{A}^0$ ever proves that it does not perform some particular action, $Axm^0 \vdash \neg\bar{b}_i$. Any conclusion would then follow from the contradictory hypothesis $\bar{b}_i$, thus $Axm^0 \vdash \bar{b}_i \rightarrow (\emptyset \vee \mathcal{G})$. With valid reasoning on true axioms this should never occur (we should not prove $\neg\bar{b}_i$ and then do b_i) but it is potentially a concern when trying to build systems which bypass the Löbian obstacle. It is ordinarily Gödel's Theorem which prevents a system from concluding that it will never prove some theorem (such as the theorem $\bar{b}_i \rightarrow (\emptyset \vee \mathcal{G})$).

torization for causal models (Pearl 2000):

$$P(\mathbf{X}) = \prod_{x_l \in \mathbf{X}} P(x_l \mid \mathbf{pa_l}), \tag{7.3}$$

where for each variable x_l, $\mathbf{pa_l}$ is a possibly empty list of the immediate causal ancestors of x_l. Thus the joint probability distribution $P(\mathbf{X})$ factorizes into local rules governing the probability of each node x_l given its parent causes, if any.

We then axiomatize $\mathcal{A}^1$'s knowledge of $P(\mathbf{X})$ in a way that leaves some parentless x_l with unspecified probabilities; any causal descendants of these variables will have probabilities unquantified by $\mathcal{A}^1$. One of these variables will be $\mathcal{A}^0$'s action and thus $\mathcal{A}^1$ will not precompute a probability of $\mathcal{A}^0$'s action. We will also need to leave unspecified all sensor variables $s_k \in S$. Collectively the sensor variables and the motor action variable will constitute a kind of Cartesian boundary between the deterministic world in which $\mathcal{A}^0$'s cognition occurs, and the probabilistic world it wants to sense and affect.

We then mate deterministic reasoning to probabilistic reasoning via the bridging axioms:[51]

$$\mathcal{A}^1 \Vdash \forall s_k \in S : \bar{s}_k \rightarrow (P(\bar{s}_k) = 1) \tag{7.4}$$

$$\mathcal{A}^1 \Vdash \forall b_i \in Acts^0 : \bar{b}_i \rightarrow (P(\bar{b}_i) = 1). \tag{7.5}$$

$\mathcal{A}^1$ and $\mathcal{A}^0$'s criteria of action will be:

$$\bar{\mathcal{A}}^0_j \Rightarrow \mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0_j \rightarrow (\emptyset \vee (\mathbb{E}[U] > \$20)), \tag{7.6}$$

$$\bar{b}_i \Rightarrow \mathcal{A}^0 \Vdash \bar{b}_i \rightarrow (\emptyset \vee (\mathbb{E}[U] > \$20)), \tag{7.7}$$

and let $\mathcal{A}^1$ prove a soundness schema (central trust lemma) over $\mathcal{A}^0$'s logical beliefs as before. Then $\mathcal{A}^1$'s decision to create $\mathcal{A}^0$ is straightforward via:

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow \forall b_i : \bar{b}_i \rightarrow (\Box_0 \ulcorner \bar{b}_i \rightarrow (\emptyset \vee \mathbb{E}[U] > \$20) \urcorner), \tag{7.8}$$

$$\mathcal{A}^1 \Vdash \Box_0 \ulcorner \bar{b}_i \rightarrow (\emptyset \vee \mathbb{E}[U] > \$20) \urcorner) \rightarrow (\bar{b}_i \rightarrow (\emptyset \vee \mathbb{E}[U] > \$20)), \tag{7.9}$$

$$\mathcal{A}^1 \Vdash (\forall b_i : \neg \bar{b}_i) \rightarrow \emptyset, \tag{7.10}$$

$$\mathcal{A}^1 \Vdash \bar{\mathcal{A}}^0 \rightarrow (\emptyset \vee \mathbb{E}[U] > \$20). \tag{7.11}$$

Suppose Axm^0 incorporates standard probability axioms, and suppose that $\mathcal{A}^0$ has observed all available sensor data s_k representing all missing terms of $\mathbf{X}$ except its action b_i. Then $\mathcal{A}^0$ will be able to carry out an action with

51. The form of these axioms was suggested by Paul Christiano.

$\mathbb{E}[U \mid \bar{s}_k \wedge \bar{b}_i] > \$20$, via:

$$\mathcal{A}^1 \Vdash \mathbb{E}[U \mid \bar{s}_k \wedge \bar{b}_i] > \$20 \tag{7.12}$$

$$\mathcal{A}^1 \Vdash \bar{s}_k \tag{7.13}$$

$$\mathcal{A}^1 \Vdash P(\bar{s}_k) = 1 \tag{7.14}$$

$$\mathcal{A}^1 \Vdash \bar{b}_i \to (P(\bar{b}_i) = 1) \tag{7.15}$$

$$\mathcal{A}^1 \Vdash (P(\bar{b}_i) = 1 \wedge P(\bar{s}_k) = 1) \to (P(\bar{b}_i \wedge \bar{s}_k) = 1) \tag{7.16}$$

$$\mathcal{A}^1 \Vdash \bar{b}_i \to (P(\bar{b}_i \wedge \bar{s}_k) = 1) \tag{7.17}$$

$$\mathcal{A}^1 \Vdash \bar{b}_i \to (P(\neg(\bar{b}_i \wedge \bar{s}_k)) = 0) \tag{7.18}$$

$$\mathcal{A}^1 \Vdash \mathbb{E}[U] = (\mathbb{E}[U \mid \bar{b}_i \wedge \bar{s}_k]P(\bar{b}_i \wedge \bar{s}_k)) + (\mathbb{E}[U \mid \neg(\bar{b}_i \wedge \bar{s}_k)]P(\neg(\bar{b}_i \wedge \bar{s}_k))) \tag{7.19}$$

$$\mathcal{A}^1 \Vdash \bar{b}_i \to (\mathbb{E}[U] = \mathbb{E}[U \mid \bar{b}_i \wedge \bar{s}_k]) \tag{7.20}$$

$$\mathcal{A}^1 \Vdash \bar{b}_i \to (\mathbb{E}[U] > \$20) \tag{7.21}$$

□

We confess this to be a moderately grotesque hack that fails almost entirely to rise to the challenge of non-monotonic probabilistic reasoning. As remarked, we included it only to serve as something that could be improved upon. The argument above goes through only because, by assumption, every possible variable correlated with U has been observed and so no further probabilistic knowledge can be gained to invalidate the conclusion.

It also is not immediately obvious to us how to extend the above hack to sensory observation of variables with causal parents. In usual practice, sensors observe the causal descendants of variables of interest—e.g., the light reflecting from an object takes on a pattern caused by that object; the observed photons are causal descendants of the event of their reflection. Being unable to represent the standard idiom of evidence in Bayesian networks is a very severe defect.[52]

Similarly it seems to us that it should be possible to compute the probability of evidence before observing it, since this practice is central to Bayesian probability. In the formalism as it stands, being able to assign a prior probability to any evidential variable would produce a logical contradiction after a sensory event added the conclusion that the probability was 1.

We suspect that there ought to be some much more general schema of trust in probabilistic soundness of a successor agent, possibly invoking explicit reasoning about probability measures in the same way that we previously introduced set theory as a backdrop to general reasoning about semantic truth, or using some probabilistic trust schema such as possibly:

$$P_\alpha(\phi \wedge (\mathcal{A}^\beta \Vdash \ulcorner P_\beta(\phi) = p \urcorner)) = p \cdot P_\alpha(\mathcal{A}^\beta \Vdash \ulcorner P_\beta(\phi) = p \urcorner). \tag{7.22}$$

Such work remains in progress, however, and in general the problem of self-modification in probabilistic agents remains wide open.

52. We suspect that this desideratum might be satisfiable by some very simple technical method we have not yet invented as of this report.

8 Bounded agents: Maximizing, satisficing, and meliorizing

Expected utility is usually discussed in the context of expected utility maximization—choosing from among the actions in a set the action with highest expected utility. We did not attempt to frame our probabilistic agent that way, because introducing the notion of *maximization* creates new challenges to tiling, separate from the challenges of probability. I.e., suppose that we are in a deterministic environment, creating offspring which will have various possible sensors, various possible ranges of action, and whose actions (given the value of background variables) deterministically score various utilities. We could possibly adapt the logical goal $\mathcal{G}$ to state some version of the highest possible score being achieved, with the successor otherwise allowing a $\emptyset$ outcome. However, to properly rise to the challenge of maximization would imply creating the successor $\mathcal{A}_j^0$ which, of those options $Acts^1$ available to you, would lead to the best score you could've received, after $\mathcal{A}_j^0$ tries to maximize its own score. It was not obvious to us how to make the proof of such a tiling go through. It is also not obvious to us that this is the correct problem on which to spend effort, since, in real-world environments, achieving the *maximum possible* expected utility is impossible for all but the simplest problems.

Maximization in full generality requires correctly assessing the score (expected utility) of every action in order to guarantee that we choose the best; this is what renders, e.g., optimal chess intractable. Satisficing is more realistic for bounded agents, in this sense, because we only need to evaluate the expected utility of a single action in order to have our decision criterion $\mathbb{E}[U] > 20$ go through. (Likewise, establishing some partially uncertain, probabilistic expectation of utility is more realistic than correctly predicting an exact outcome or score.) But to seriously consider satisficing as the foundation of a useful rational agent, or sufficiently advanced self-improving intelligence, would come with its own set of problems. Suppose the lower bound is set at $\$2 \cdot 10^{15}$ utility in whatever measure, and the agent realizes that a certain plan will ensure a minimum expected utility of $\$10^{16}$. Then as we have defined the agent's goals, it satisfices to randomize between a 30% probability of implementing this plan, a 60% probability of doing something which leads to expected utility \$0, and a 10% probability of doing something with expected disutility $-\$10^{16}$. This is probably not the sort of behavior we wanted.[53]

Schmidhuber's "Gödel machine" (2007)[54] introduced the interesting suggestion of a *swapping* criterion of action: The "Gödel machine" would change its own source code if it could prove that the action of changing its source code had a higher expected utilty than keeping its current source code. (No criterion for external motor actions was introduced, but *arguendo* the Gödel machine would soon choose to add one.) One of the participants at the April 2013 MIRI work-

53. You get what you pay for, no more than you asked for, and exactly what you proved.

54. Schmidhuber (2007) could easily be considered a spiritual forerunner of the present work.

shop suggested dubbing this a "meliorizing" criterion of action, since it is meant to make things steadily better (repeatedly swap in plans, or successors, with higher expected utility). Furthermore, since meliorizing requires considering only two courses of action—the current plan, and the new plan—it may be suitable for a boundedly rational agent.

In contrast to the suggestion in section 4 of Schmidhuber (2007) it is not obvious that the criterion, "Swap if expected utility of the new program is higher" is a guarantee of "global optimality," which Schmidhuber suggested would be the result of each considered swap taking into account the possibility of other, better swaps if the current program were left undisturbed. Even considering expected utility rather than utility, most expectations that can be computed over some series of branches will not locate the point of highest expected value in the space, unless expectations are unusually smooth or the suggested series of changes is especially good. E.g., the problem at hand could easily have an NP-hard optimal solution (while still having good non-optimal solutions which could be steadily improved). But the question of "global optimality" is probably not the most important concern, since literal global optimality in the sense of trying to solve NP-hard problems should not be the key research desideratum.

It is likewise not obvious to us that "meliorizing" is sufficient to produce satisfactory behavior with respect to a builder's programmed set of goals. Suppose a sufficiently advanced machine intelligence, built according to this criterion, discovered that an asteroid was headed toward Earth and would shortly kill 7 billion people, with its current plan not preventing it. Under the strict criterion of meliorizing as written, it would make sense to swap to a program that promised to save 1,000 people, let all the others die, and make no further improvements, since this would still be better than not swapping. According to the line of argument in section 4 of Schmidhuber (2007), the agent ought to consider that it would be better to keep the previous program and wait for it to generate a better alternative. But this relies on some particular sequence of suggestions being generated such that a better alternative is considered at some point; moreover, that the agent probabilistically expects that such a better alternative will be generated if it keeps its current program, in advance of considering the actual alternative (which the Vingean principle says we cannot always do at the earlier decision point). Thus if a meliorizing criterion is ultimately satisfactory, it will be due to other properties of the series of suggestions being considered, and the way in which expectations of old and new programs are evaluated, which have not been specified into the "swapping" rule itself.

But expected utility satisficing is not satisfactory at all, and is probably not repairable; and maximizing is only possible for the best imaginable agents, not the agents that will actually exist; whereas it might be that meliorizing can somehow be improved upon. More generally, we have discussed maximizing, satisficing, and meliorizing in order to make the point that when it comes to bounded, probabilistic rational agents that are meant to pursue their goals in some "reasonable" way and build descendants who do the same (equivalently self-improve), we are not able to presently state—even on the highest possible

level of generality such as "satisficing" or "meliorizing"—what sort of criterion of action might be suitable. The problem is very wide open indeed.

References

Altair, Alex. 2013. "A Comparison of Decision Algorithms on Newcomblike Problems." http://intelligence.org/files/Comparison.pdf.

Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In "Theory and Philosophy of AI," edited by Vincent C. Müller. Special issue, *Minds and Machines* 22 (2): 71–85. doi:10.1007/s11023-012-9281-3.

Christiano, Paul F., Eliezer Yudkowsky, Marcello Herreshoff, and Mihaly Barasz. 2013. "Definability of Truth in Probabilistic Logic." Working Paper, April 2. http://intelligence.org/wp-content/uploads/2013/03/Christiano-et-al-Naturalistic-reflection-early-draft.pdf.

Fallenstein, Benja. 2012. "An Angle of Attack on Open Problem #1." *Less Wrong* (blog), August 18. http://lesswrong.com/lw/e4e/an_angle_of_attack_on_open_problem_1/.

Gentzen, Gerhard. (1936) 1969. "The Consistency of Elementary Number Theory [Die Widerspruchsfreiheit der reinen Zahlentheorie]." In *The Collected Papers of Gerhard Gentzen,* edited by M. E. Szabo, 132–213. Studies in Logic and the Foundations of Mathematics. Amsterdam: North-Holland.

Girard, Jean-Yves. 1971. "Une Extension de l'Interpretation de Gödel à l'Analyse, et son Application à l'Élimination des Coupures dans l'Analyse et la Théorie des Types." In *Proceedings of the Second Scandinavian Logic Symposium,* edited by Jens E. Fenstad, 63–92. Vol. 63. Studies in Logic and the Foundations of Mathematics. Amsterdam: North-Holland. doi:10.1016/S0049-237X(08)70843-7.

Goodstein, R. L. 1944. "On the Restricted Ordinal Theorem." *Journal of Symbolic Logic* 9 (2): 33–41. http://www.jstor.org/stable/2268019.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science.* Edited by G. Larry Bretthorst. New York: Cambridge University Press. doi:10.2277/0521592712.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory.* Cambridge Studies in Probability, Induction and Decision Theory. New York: Cambridge University Press. doi:10.1017/CBO9780511498497.

Kaivola, Roope, Rajnish Ghughal, Naren Narasimhan, Amber Telfer, Jesse Whittemore, Sudhindra Pandav, Anna Slobodová, et al. 2009. "Replacing Testing with Formal Verification in Intel ® Core™ i7 Processor Execution Engine Validation." In *Computer Aided Verification: 21st International Conference, CAV 2009, Grenoble, France, June 26 – July 2, 2009. Proceedings,* edited by Ahmed Bouajjani and Oded Maler, 414–429. Vol. 5643. Lecture Notes in Computer Science. Springer. doi:10.1007/978-3-642-02658-4_32.

Kirby, L., and J. Paris. 1982. "Accessible Independence Results for Peano Arithmetic." *Bulletin of the London Mathematical Society* 14 (4): 285–293. doi:10.1112/blms/14.4.285.

Kruskal, J. B. 1960. "Well-quasi-ordering, the Tree Theorem, and Vazsonyi's Conjecture." *Transactions of the American Mathematical Society* 95 (2): 210–225. doi:10.1090/S0002-9947-1960-0111704-1.

LaVictoire, Patrick, Mihaly Barasz, Paul F. Christiano, Benja Fallenstein, Marcello Herreshoff, and Eliezer Yudkowsky. 2013. "Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic." Preprint. http://intelligence.org/files/RobustCooperation.pdf.

Löb, M. H. 1955. "Solution of a Problem of Leon Henkin." *Journal of Symbolic Logic* 20 (2): 115–118. http://www.jstor.org/stable/2266895.

Newell, Allen, J. C. Shaw, and Herbert A. Simon. 1959. "Report on a General Problem-Solving Program: Proceedings of the International Conference on Information Processing." In *Information Processing,* 256–264. Paris: UNESCO.

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

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference.* 1st ed. New York: Cambridge University Press.

Schmidhuber, Jürgen. 2007. "Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers." In *Artificial General Intelligence,* edited by Ben Goertzel and Cassio Pennachin, 199–226. Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-540-68677-4_7.

Strachey, Christopher. 1967. "Fundamental Concepts in Programming Languages." Unpublished notes from International Summer School on Programming Languages, Copenhagen. Reprint. 2000, *Higher-Order and Symbolic Computation* 13 (1–2): 11–49. doi:10.1023/A:1010000313106.

Tarski, Alfred. (1935) 1983. "The Concept of Truth in Formalized Languages [Der Wahrheitsbegriff in den Formalisierten Sprachen]." In *Logic, Semantics, Metamathematics: Papers from 1923 to 1938,* 2nd ed., edited and translated by J. H. Woodger, 152–278. Indianapolis: Hackett.

Vinge, Vernor. 1984. *True Names.* New York: Bluejay.

Weaver, Nik. 2005. "Predicativity Beyond Gamma$_0$." Unpublished manuscript. Last revised May 11, 2009. http://arxiv.org/abs/math/0509244.

Willard, Dan E. 2001. "Self-Verifying Axiom Systems, the Incompleteness Theorem and Related Reflection Principles." *Journal of Symbolic Logic* 66 (2): 536–596. http://www.jstor.org/stable/2695030.

Yudkowsky, Eliezer. 2011. "Open Problems in Friendly Artificial Intelligence." Paper presented at Singularity Summit 2011, New York, October 15–16. http://www.youtube.com/watch?v=MwriJqBZyoM.