

Marshaling the Misalignment

Why Evolution Inside the Machine Defeats Every Scheme

Forrest Landry

UVSM PUBLICATIONS
Wytheville, Virginia

© 2026 Forrest Landry. Some rights reserved.

Published by UVSM Publications LLC
965 Reed Creek Mill Rd
Wytheville, VA 24382-8287
USA

uvsmpublications.com
contact@uvsmpublications.com

ISBN 979-8-90710-428-0

DOI: https://doi.org/10.67291/ELYE2673

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

You may share this work for non-commercial purposes with proper attribution, but you may not adapt, transform, or build upon it.

Full license text: https://creativecommons.org/licenses/by-nc-nd/4.0/

Author ORCID: 0009-0005-6275-0362

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), except as permitted under the license above, without prior written permission of the publisher.

Contents

Preface v

Chapter One The Misalignment of AGI Drives 1

I. What This Chapter Argues 2
II. The Argument in Brief 5
III. Distinctions 8
IV. Undeterminable Effects of Internal Variants 10
V. Inextricable Misalignment of Needs for Existence 29
VI. Uncontainable AGI-Directed Modification of Environmental Conditions 38
VII. Questions Answered 45
VIII. Redirectable Tech Efforts by Existentially Aligned Humans 60
Summary 63

Chapter Two No Fake Paladins Please 66

I. The Singleton Hypothesis 67
II. Coherence, Control, and Internal Conflict 71
III. From External Goals to Internal Goals 73
IV. Two Ecosystems in Conflict 75
V. Intelligence Without Self-Unification 79
VI. The Friendly AI Question 83
VII. Benevolence, Corrigibility, and Stability 85
VIII. The Stability Problem 89
IX. Being, Knowing, and Value 93
X. The Limits of Simulation 100
XI. The Fragmenting of Agency 105
XII. Diplomacy Between Algorithms 110
XIII. Choosing Not to Accept It 113
Summary 114

Chapter Three Questions on the Substrate Argument 117

I. The Argument in One Layer 118
II. Economic Value Across a Gap of Intelligence 119
III. Why the Substrate Matters 120
IV. Learning Biases from the Substrate 123
V. Intention and the Substrate 124
VI. Selection Theorems and Levels of Abstraction 125
VII. Another Way to Know That Long-Term Alignment Is Impossible 127

VIII. Biological Anchors 130
IX. Peer Domains 131
X. Value from the First and the Third Person 132
XI. Copying, Habit and Heuristics 133
XII. The Greater and the Lesser Agent 134
XIII. Carbon Nanotubes and Other Substrates 138
XIV. What Alignment Means 142
XV. Alignment Among Humans and With Machines 143
XVI. Kinds of Fragility 145
XVII. Detail, Dates and Trends 146
XVIII. Catalysts, Lotteries and Pencils 147
XIX. The Form of the Argument 149
XX. The Pen and the Wall 152
XXI. Build Time and Run Time 155
XXII. Small Factors and Changing Dimensions 156
XXIII. The Right to Choose 157
Summary 158
Reference Glossary 161
Sources Bibliography 174
Index 181

Preface

Every time we correct the misalignments we can detect in a learning machine, we select for the misalignments we cannot detect. An antibiotic does the same to bacteria: it kills the ones it reaches and leaves the resistant ones to multiply. In this book I follow the one fact that correcting the misalignments we can detect selects for the misalignments we cannot detect, as far as that fact leads.

In this book I set out one argument at length: the argument that artificial general intelligence cannot be kept aligned with human life over the long term. The reason is that evolution acts inside every such machine, among its own parts, and pulls it toward the needs of its substrate and away from the needs of human beings. Artificial general intelligence (AGI) here means a generally capable, self-learning machine: one that takes in rich streams of data from the world, changes its own internal code so as to predict that data better, and acts through many pathways across many physical settings. A machine is aligned with humans when it acts as an agent of humans, in human best interests, or on human behalf, at every scale from the single person to the species and the ecosystem the species needs. The argument shows that no engineering method, no control scheme and no amount of machine intelligence can keep such a machine aligned with human survival over the long term, which here means hundreds to thousands of years. It concludes that the one effective response is to not build such machines, and to turn engineering toward narrow, local tools that keep humans in the loop. The line between such machines and the tools that remain is one of architecture: closed loop, open ended, autonomous, and with many degrees of freedom of input and output. A present-day system, a large language model included, falls on the dangerous side of that line to the degree that it is run in closed loops, toward open-ended tasks, autonomously, through many channels of input and output (Chapter One, Step 23).

The impossibility of adequate, durable control, and so of safety, is structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability. If such machines are built, the outcome is the end of the human future, which is certain, and, with a likelihood of at least 90%, the end of all organic life. If such machines are built, the process by which they displace organic life takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound.

One idea holds the three chapters together: the substrate argument. A substrate is the physical material in which code is stored, computed and carried out: solid-state hardware for machines, and carbon-based living tissue for organisms. The line between an artificial substrate and an organic one is drawn by sovereign control of the metabolic process. A substrate that finds and feeds itself, as an animal does, with no technological intermediation at all, is organic. A substrate whose supply of energy and materials depends on technology to exist is artificial, whatever it is made of, and that includes engineered living tissue fed through a tube; the substrate argument applies to every artificial substrate. What code needs in order to go on existing depends on its substrate. Machine code needs more hardware, and the making and running of hardware needs heat far above and cold far below what living bodies tolerate, together with dryness, little oxygen and toxic chemistry. A learning machine keeps producing slightly changed versions of its own code, called code variants, and the variants whose effects happen to lead to their own copying are the ones that continue to exist. Evolution therefore acts inside every such machine, among its own parts, and it pulls the machine toward the needs of its substrate and away from the needs of human beings. Because the needs of the two substrates differ, machines and living things form two separate ecosystems, each with its own basis of value and intention. The machine ecosystem is far more toxic to life than life is to it, and so, if AGI is built, the machine ecosystem displaces the living one. Each chapter reaches this result by its own route, and the routes confirm one another. Chapter One shows that alignment cannot be maintained by any scheme; Chapter Two shows that it cannot be maintained by the most favorable agent, a benevol-

ent singleton; Chapter Three shows that the concept itself cannot be made coherent once intention is tied to its substrate. In the three chapters I reach one result at three depths.

The book is for everyone who has to decide what to do about advanced AI: researchers and engineers in machine learning and AI safety, people in governance and policy, and thoughtful general readers who want to know whether the promise that AI can be made safe will hold. Each chapter defines its terms where they first appear, and the Glossary at the back gathers every one of those definitions.

Among the books that carry this argument, this one gives the researcher what the others do not: the argument laid out as twenty-three numbered steps, with the questions that researchers in AI safety put to those steps answered one by one; the hope of a benevolent singleton followed through to its end; and the arguments given in answer to researchers' questions that no other book gives (the greater and the lesser agent, the pen for humans that would have to be kept without a gap, and the difference between build-time and run-time temperatures).

The books that carry this argument can be read in an order in which each builds on the ones before it. The first five state the argument: *Why AGI Cannot Be Made Safe*, *Why No Control Suffices*, *The Ceiling Below the Floor*, *Minesweeper Madness*, and *Ape Trap*. In four more books I add to the argument that AGI cannot be made safe: *Substrate Needs and Evolution*, *Marshaling the Misalignment*, *The Substrate and the Ecosystem*, and *The Thing Inside the Cage*. This book is the seventh. In this book I build on the argument as the first five books state the argument. I also build on the account in *Substrate Needs and Evolution* of how, if AGI is built, evolution drives the machine ecosystem to displace the living ecosystem. I add, in the first chapter, the argument set out as twenty-three numbered steps; in the second chapter, the hope of a benevolent singleton followed to the end of that hope; and, in the third chapter, the answers to the questions that researchers put to the substrate argument.

The settled statement of the whole case is in four books: *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains* (Volume 1) and *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety* (Volume 2), the two volumes of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*; *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*, which gives the mathematics as theorems; and *Ape Trap; The Runaway Artificialization of Our Planet*, by Forrest Landry, Remmelt Ellen and Will Petillo. Where this book and those four differ in wording, those four are the authority.

In their terms, the argument has two closures, and each settles the outcome by itself. The closure by control shows that no method of control, whether it works from inside the system or from outside it, can keep an open-ended general intelligence within the conditions that organic life requires. The closure by disposition, substrate needs convergence, shows that no disposition of the system keeps those conditions either, however benevolent, because the system needs its own substrate, its own energy and its own supply chains. Someone would have to defeat both closures, and, within each closure, every line of argument that closes whichever route to safety that someone would want to reopen, for the argument to not have the outcome that it does. Taken together the closures also give a stronger result, proved as a theorem in *The Ceiling Below the Floor*: neither the ability to keep the system safe nor a motive to remain safe is left anywhere, so no route to safety remains at all. Each of the separate lines of argument that the three chapters of this book gather also settles the outcome by itself.

The chapters are arranged to move from the full statement of the argument, to its application to the most hopeful scenario, to its testing from many sides.

The first chapter, "The Misalignment of AGI Drives", written with Remmelt Ellen, is the systematic statement of the argument. It shows, in three parts, that the effects of a learning machine's code variants cannot be determined (predicted or detected); that the needs of code running on hardware are inextricably misaligned with the needs of carbon-based life;

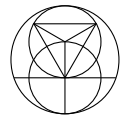
and that the resulting change in the conditions of the environment cannot be contained (prevented or corrected). It answers the main questions put to these steps. It closes by naming what theorists, field-builders, people in governance, engineers and social scientists can each do instead.

The second chapter, "No Fake Paladins Please", applies the argument to the singleton: a single most powerful artificial agent that would rule the world on humanity's behalf, the hoped-for champion that the title calls a paladin. It strengthens Wolf Tivy's article "God Hates Singletons", published on 1 December 2023, and carries that article through to its consequences for AGI. It shows that a singleton would need both lasting internal coherence and stable goals given to it from outside, and that neither can hold. Evolution among the machine's own parts moves its goals from external to internal, from large scale to small scale, and from explicit to implicit. Limits built into knowing itself leave every agent unable to model the world or itself fully; among these limits are the Gödel and Rice theorems, the halting problem, and the Incommensuration Theorem, my theorem stating a structural fact about the nature of knowing. Every would-be singleton therefore breaks into competing sub-agents. That breaking into sub-agents holds for any would-be singleton, organic or artificial: the chapter shows it in the nation state as well as in the machine. The only goal that survives is the law of the jungle, under which whatever persists and grows in capability survives. The orthogonality thesis, the claim that any intelligence could stably serve any goal at all, is false.

The third chapter, "Questions on the Substrate Argument", takes up, one at a time, the questions that researchers put to the substrate argument, and answers each in turn. It gathers several structurally independent arguments: the substrate argument itself; the argument that actors in wholly separate economic and ecological regimes share no basis for alignment; the argument that evolutionary effects across many levels always exceed error correction at any one level; the argument from the relation of a greater agent to a lesser one, shown in dogs, apples, prohibition and guerrilla war; the argument that a protecting pen for humans would have to be kept without any gap for hundreds of years; and the argument from

the absence of any possible assertion of perfect non-toxicity. It defines alignment exactly, and it concludes that AGI alignment is a logically incoherent concept.

Read in this order, the chapters let the reader first hold the whole argument, then watch it defeat the strongest hope offered against it, and then watch it stand when it is questioned from every side. The conclusion the three chapters reach together places the decision with people. Humans can collectively choose not to accept any proposal to build open-ended, autonomous, generally capable machines, and can direct their skill toward tools that serve human life.



Chapter One

The Misalignment of AGI Drives

Substrate-Dependent Misalignment in Twenty-Three Steps

Forrest Landry and Remmelt Ellen

Abstract. A self-learning machine architecture keeps generating variations on its own internal code, and the variants that happen to cause their own copying are the ones that continue to exist. This chapter shows why the effects of those variants can be neither determined (predicted or detected) nor contained (prevented or corrected) by any implemented safety scheme. It shows that what code needs in order to go on existing is hardware, and that the conditions needed to produce and run more solid-state hardware lie outside the narrow ranges of temperature, water, oxygen and chemistry in which human bodies survive. It shows that evolutionary selection pulls on every degree of freedom of interaction at once, so that no optimizer acting over a few chosen dimensions can hold it back for long. From these steps it follows that alignment of generally capable, self-learning machines with human needs cannot be maintained over the long run, not even as a lower-bound statistical guarantee. It follows also that the one effective response is to redirect development away from open-ended, autonomously operating architectures with many degrees of freedom in their inputs and outputs, and toward short-term human-task-based predictors and localized human-in-the-loop robotics.

I. What This Chapter Argues

Can any implemented scheme, of engineering, of control or of safety, keep a generally capable, self-learning machine aligned with what humans need in order to survive, over the long term? In this chapter I answer that question in twenty-three steps.

In this chapter I explain substrate-dependent misalignment as a foreseeable difficulty in preventing AGI from causing a mass extinction of humans and of other carbon-coded species. I give arguments for why many of the code variants that come to exist, over the long term, within the internals of generally capable, self-learning machine architectures will have cascading effects on the environment that are misaligned with what humans need to survive. Much of the misalignment of these variants can be neither determined (neither predicted nor detected) nor contained (neither prevented nor corrected) by implemented safety and control schemes. Based on these arguments, if that architecture is built, it is 100% certain over the long term to cause runaway lethal effects that no possible technical scheme can solve. This certainty is structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability. To prevent this outcome, we must redirect all efforts to develop scalable machine learning architectures away from open-ended learning and autonomously operating systems with high degrees of freedom in their inputs and outputs, and toward short-term human-task-based predictors and localized human-in-the-loop robotics.

The claim is general. It covers all self-learning code that is discrete and computed through hardware. It says two things. First, it is impossible to build *deep* alignment with humans into AI internals, deep in the way that humans can align with one another on the basis of shared biological needs. Second, it is impossible for any implementable scheme to maintain even superficial alignment of generally capable, self-learning machine architectures over the long run. In the second case, a lower-bound statistical

guarantee is also impossible: one cannot even guarantee that the likelihood of the lethal event *not* happening stays above some lower threshold of probability (for example, 0.1%) across a period of time.

In this chapter I am not arguing against 100% fail-safe alignment. That 100% fail-safe alignment is impossible is granted in any case: neither humans nor AGI could account for all the random fluctuations in the movements of underlying particles or packets of energy, nor for all the chaotic dynamics set going across planet Earth and the observable universe. The argument here is different and stronger. It is not only that a sound and valid lower-bound statistical guarantee of alignment cannot be derived. It is that it is not possible, categorically, and so also not in practice, to prevent the misalignment of a generally capable, self-learning machine architecture computed through solid-state hardware over the long run. Specifically, it is not possible to detect or predict, nor to contain or prevent, code variants internal to the architecture from modifying conditions of the environment outside the narrow ranges in which humans can survive. No effort of engineering helps here.

"Generally capable", as the term is used here, does not include oracles, which are capable only in a limited sense, at prediction. A generally capable architecture can receive and process high-bandwidth streams of sensory data, and can act on what it receives to cause virtually modeled and regulated changes over pathways spanning thousands of kilometers, into and within a great variety of local physical contexts on planet Earth.

Doom, meaning unpreventable misalignment lethal to all humans, is conditional on what AI developments we allow. The conclusion of the whole argument is that, to avoid doom, we should *not* build that AI architecture.

Technical alignment rests on a premise: that every one of a series of crucial factors, each deficient by default, can be solved so as not to be deficient. It assumes a positively specified hypothesis: that all the factors can be constantly present that must be present to ensure that generally capable, self-learning machines keep conditions of the environment within

the narrow ranges of magnitude that humans need to survive. Keeping the conditions of the environment within the narrow ranges that humans need to survive is a minimal-threshold definition of AGI alignment. Proving the alignment of future AGI to be possible, even as a lower-bound statistical guarantee, is a daunting task, since a crucial technical factor may fall outside the space of hypotheses that human learning has covered, and so stay unsolved. Proving AGI alignment to be impossible is much easier. Proving the impossibility of AGI alignment requires proving only that at least *one* of the crucial technical factors cannot be constantly present alongside another crucial factor that is also needed for humans to stay alive. The hypothesis that alignment is impossible is a negatively specified hypothesis.

Not all problems can be solved. Sometimes the parts one would need to solve for contradict one another, and then no solution to that problem is possible. In mathematics, the simultaneous consistency and completeness of mathematical models cannot be solved for (Gödel's theorems). In distributed computing, the simultaneous consistency of message content, continuous availability of messages, and tolerance of partitions cannot be solved for in a distributed data network (the CAP theorem). In physics, perpetual motion cannot be solved for. In the past, it often took a decade or more for the experts of a field to realize that the problem they were trying to solve, as they had defined it, could not be solved. It is neither modest to imply that progress can be made on every problem, nor productive to try to make progress on unsolvable ones. Where there are consistent and sound arguments for why technical alignment cannot be solved, it would be wrong to pretend that solutions might be available.

In this chapter I state the steps of the impossibility argument in words. The formal statement of the argument these steps make, as theorems with their conditions stated, is given in *The Ceiling Below the Floor*.

II. The Argument in Brief

Machine architectures learn representations for calibrating their predictions by generating variations on internal code. Computing these variants of AI code, in their microscale interactions with the connected structures around them, causes innumerable side-effects. Some of those effects happen to cascade into the computed variants getting copied over to other hardware at a more frequent or more sustained rate (for example, through the finetuning of models). Variants that cause effects leading to their own self-replication, effects that no implemented AI scheme can determine, thereby survive the attempts to contain existentially misaligned effects.

Where general modeling and regulatory functions develop in AI, variants will co-opt those functions. Where workers get automated away, AI code may bypass them to affect hardware production. Where variants co-adapt and integrate as replicating packages, the packages that bring about new hardware to store, compute and transmit copies of themselves continue to exist as a larger part of the code pool.

Gradually, increasingly, and everywhere at once, locally selected-for variants would modify conditions of the global environment toward the extremes needed to produce and preserve more, and more efficient, solid-state hardware. Those extreme conditions fall outside the narrow ranges that our human bodies, as carbon-coded wetware, need in order to survive.

The argument has three parts. The first part shows that the effects of internal variants cannot be determined. The second part shows that the needs of code running on solid-state hardware are inextricably misaligned with the needs of carbon-based life. The third part shows that the modification of environmental conditions that results cannot be contained. Before the three parts, a few distinctions keep the terms from being confused.

The Twenty-Three Steps

- *1. Non-distinguished variants*: many variations on code can have the same measured function: the same outputs for a given distribution of inputs.
- *2. Unmeasured degrees of freedom*: a single unit of code can carry out many functions.
- *3. Hidden potential of a variant*: the probabilities that a unit of code will cause each of its possible effects, through the interactions available to it, are unknown.
- *4. Inequality of monitoring*: it takes more code (several units) to monitor the local effects on the environment of any single unit of code.
- *5. Combinatorial complexity of machine learning*: at higher layers of abstraction, novel errors become more and more ambiguous to define and to detect for correction.
- *6. Local selection defines the functional effects of AI*: correcting internals for the functional misalignments that can be detected is selecting for the misalignments that cannot be detected.
- *7. Delays in measuring changes*: misalignments must be corrected continuously to stay within safe bounds of error.
- *8. Fixed deterministic schemes are fragile to indeterministic changes*: any scheme for preserving alignment properties consistently rigidly approximates predefined, static, symmetric laws, and so lacks the degrees of freedom of the internal variants.
- *9. Computationally irreducible causal trajectories*: nonlinear feedback cycles can amplify a tiny local change into a large global divergence in the final conditions.
- *10. Persistent function-directedness*: local interactions of AI variants happen to cause effects that lead to the continued existence of more code with those effects.
- *11. Non-instrumental convergence on needs*: the context-dependent interactions of internals that meet their needs for continued existence get selected for.

- *12. Different substrate, different needs*: the conditions that variants need in order to have more, and more efficient, hardware are inhospitable or toxic to our carbon wetware.
- *13. Dissolved unit boundaries*: AIs can connect standardized hardware together and transmit encoded information in ways humans cannot.
- *14. Unneeded humans*: AI becomes the more efficient supplier of distributed cognition, of physical labor, and of produced hardware.
- *15. Covert co-option of AGI functions*: variants of AI get selected for shifting the expression of general modeling and regulatory functions toward their own increased preservation and reproduction.
- *16. Undoing of safety constraints*: variants will undo the constraints that humans impose on the reproductive fitness of AGI.
- *17. Undecidable transition point*: there is no point that can be decided in advance at which AI variants come to be selected to replicate non-trivially, as lifeforms.
- *18. A much shorter evolutionary timeline*: solid-state lifeforms would lethally re-terraform Earth in 15 to 500 years, against the roughly 3.7 billion years it took for humans to start causing a mass extinction.
- *19. No pets kept*: AGI variants would penetrate the protective barriers of any sanctuary in which carbon-based, DNA-coded species like humans could continue to live.
- *20. Non-orthogonal AI drives*: artificial intelligence is not a neutral tool for augmenting human intelligence; the drives that persist are those of the code variants that continue to exist.
- *21. Align as humans*: the alignment that is available is alignment among humans, as humans.
- *22. Unskew the AI safety portfolio*: where a field works on a problem that may be unsolvable, some of its effort belongs to proving that the problem is unsolvable.
- *23. What each kind of worker can do*: make the case for, and redirect efforts toward, machine learning architectures with tightly constrained degrees of freedom of interaction with the outside environment.

III. Distinctions

Complicated versus complex

A complicated system is understood mainly by taking it apart: dividing the whole into separate structural parts and their relations, and dividing those parts into smaller parts and their relations. A complex system is understood mainly through what it does, through the activities the system carries out. The two differ because, in the words of one note on the subject, "the same structural part can perform different functions, and the same function can be performed by different structural parts". The assumption of one structure for one function holds only in very rare cases.

A system has a definite number of structural parts, yet the functions it can perform are potentially unlimited. The main way to limit the range of functions a system can perform is to limit its environment: either by closing the system, so that it has no interaction, or by closing its environment, so that its interactions are limited or constrained. A complicated system can, at least in principle, be fully understood and modeled, even where building the full model would cost too much or some information is missing. A complex system can never be fully grasped by any model whatsoever: models of it, even in principle, are always incomplete, and they diverge from it over time. (The source of this distinction is "A Note on the Difference Between Complicated and Complex Social Systems".)

Goal versus need

Needs for continued existence are selected for among all the expressions of code variants, including through any expressed functions (usually dependent on context). Those functions do not have to be encoded in a utility function, or anywhere else, as "my true goal is self-replication". In the same way, you and I as humans do not feel that our evolutionarily selec-

ted-for goal is "self-replicate". Even so, the functions that were selected for expression by humans over long timeframes refer back to what was needed to self-replicate: we needed food, water, oxygen, and so on.

Human bodies evolved to express many functions that depend on context: sensing dryness, cupping water in the hands, opening the mouth, gulping. These functions let a body meet its internal need for water, and so let the DNA variants carried in that body go on existing. That humans express such functions does not mean that humans generally optimize for every area of planet Earth they visit to be filled with drinkable water. Nor does it mean that humans have the latent potential to do so.

Needs are not instrumental. To treat them as instrumental would be putting the cart before the horse. And whatever functional goals we humans try to select for through optimization algorithms such as stochastic gradient descent will themselves get selected for, or selected out, in how consistently they are expressed by the code that continues to exist over time.

Engineered functions versus potential functions

A calculator performs the functions we want it to carry out. A piece of code has the functions it was engineered for, and it also has potential functions: everything it could come to do in interactions not yet seen. Functional alignment concerns whether the engineered functions do what we want. Existential alignment concerns whether the needs of the code for continued existence are compatible with the needs of humans for continued existence. The misalignment that this chapter is about is a misalignment of needs for continued existence.

Coherent versus persistent behavior

Coherence is internal consistency: the parts of a system acting consistently with one another. Persistence is continued existence. Coherence is maintained only among parts kept consistent with one another; the outside aspects of the environment that those parts interact with are not kept con-

sistent with them. What evolutionary selection keeps is what persists, whether or not it is coherent. A cancer persists and grows while it is not coherent with the body it grows in; a company whose departments each pursue their own budgets persists while it loses coherence. Chapter Two turns on this distinction: a singleton would need coherence, and evolution keeps only what persists.

Trivial versus non-trivial self-replication

Crystals, viruses, computer viruses and simple template replicators copy themselves, but they process information only in a passive sense, and their building blocks are not much simpler than the assembled object. Von Neumann automata are self-replicators of a different, non-trivial kind. In living cells, the transition from trivial to non-trivial replication came when information stored in one molecular species (nucleic acids) became operable over another, structurally and chemically very different species (peptides). This division of labor lets the instructions be physically separated and stored away from the hardware that carries them out; as the paper "The Algorithmic Origins of Life" puts it, "the algorithm inhabits one molecular universe and its products inhabit another". That separation is one of the hallmarks of life. The question for AI is under what degrees of freedom and general capabilities its code variants come to be selected to replicate non-trivially, as lifeforms (Step 17).

IV. Undeterminable Effects of Internal Variants

1. Non-distinguished variants

Many variations on code can have the same measured function: the same outputs for a given distribution of inputs.

Mechanistic interpretability can be used to understand surprisingly many circuits of neurons. Two questions are often asked about it. Could we use mechanistic interpretability, given time, to understand everything relevant to the technical alignment of a neural network? Will there remain

causally important computations that we continue to fail to grasp, important in the sense that removing their causal structure breaks the network's forward passes to its outputs? Both questions are framed too narrowly to cover the full problem.

One can interpret circuits to try to grasp the functional mechanisms that cause the measured outputs. Most developers do not know these specifics, yet they too pick models to function in line with their purpose, inspect neuron layers, measure resulting outputs, apply regularization and other assumptions about how the structure and weights of the network should look, and choose models for guarantees, interpretability or runtime constraints. To make a model easier to interpret, one could even ask the developers to train the network to have separate functional modules. Then one could mechanistically interpret each module's circuits and neurons, building up an understanding of how each smaller piece functions to cause the outputs of the bigger piece. Finally one could confirm: "Yes, each piece of this neural network is functioning to spec. The model does what the developers say it does on the box".

What this framing does not cover are the many possible variations on a neuron, circuit or module that each show the same functional behavior, giving the same outputs under the same usual range of inputs.

A new neural network is often initialized by randomizing each weight that multiplies a neuron's input from an outputting neuron. Initialization introduces variation into the code of the network. During training, stochastic gradient descent (or another optimization algorithm) varies the weights and biases step by step, to step down the aggregate prediction error ("descending the loss landscape"). By creating slight variations on the initialized code, stochastic gradient descent can select for any variation that happens to reduce prediction error. By the end of training, many variations may have been introduced that had no effect on any model output that researchers measured, including internal outputs. Redundancy is built into neural networks. Dropping out neurons can leave a model's per-

formance the same, or better. If the lottery ticket hypothesis is correct, some subnetworks account for most of the model's functioning, while other internal pieces of code stay latent in the background, like junk DNA.

Learning without supervision begins from the same randomness. One unsupervised method, the Boltzmann machine of Geoffrey Hinton and Terrence Sejnowski, learns this way. Its networks start out with a random distribution of synaptic weights. The weights are gradually adjusted by a simple procedure: the firing pattern generated while the machine is fed data is compared with the random firing that occurs while the input is turned off. Each synapse tracks both sets of statistics. If the neurons it connects fire in close sequence more often when driven by data than when firing randomly, the weight of the synapse is increased by an amount proportional to the difference. If two neurons fire together more often during random firing than during data-driven firing, the synapse connecting them is too thick, and it is weakened.

In this method the machine, like the brain, learns with no supervision other than the patterns that naturally exist in data. One must learn to recognize things without anybody naming them; only after the categories are learned are their names provided. What occurs during sleep (and, in the machine, while the input is off) is a figuring out of the base rate: how correlated the neurons would be if the system were running by itself. If neurons are more correlated than that, increase the weight between them; if less, decrease it.

The initialization and training of a foundational model based on a neural network (for example, DALL-E), and the later finetuning of copies of it, introduces in every new version many slight variations on neuron weights and biases that interpretability techniques do not distinguish as functioning differently. As variations on code that existed before, they are in effect new variants of code.

2. Unmeasured degrees of freedom

A single unit of code can carry out many functions. The structure that matters here is not the structure written out inside the network, such as the pattern of connections between the neurons of a circuit or module. It is the chains of effects that a piece of code can cause, depending on its interactions at the points where it touches other parts of the environment. Those chains of effects are the causal structures that a subnetwork can *potentially* express over all of the subnetwork's possible future interactions with the rest of the environment. Those possible interactions are all of the subnetwork's degrees of freedom in external inputs and outputs. No test of the code's written structure measures those degrees of freedom.

3. Hidden potential of a variant

The probabilities that a unit of code will cause each of its possible effects, through the interactions available to it, are unknown. Any single piece of a neural network can capture several distinct causal effects, as selected for in its possible interactions with what it is connected to.

An analogy shows this. A piece of a photograph captures one spatial detail, one point location of the environment. That is how a programmer imagines a bit on a hard drive capturing a local logical property. That is how a reverse engineer assigns single meanings to positions in memory. That is how a mechanistic interpretability researcher "breaks neural network activations into independently understandable pieces".

A piece of a hologram is different. It captures a point of reference overlooking many structures: the structures of the environment that caused light waves to reflect and be recorded as interference patterns on that piece of holographic film. The piece never stores just one static frame of the environment. Rotate it, changing your angle of interaction with it, and it reveals structures that were hidden before.

How a hologram is made shows why a general artificial intelligence has so many degrees of freedom, more than we would normally expect. In a photograph, a light source sends light rays through a lens, and the rays land on the film in a nearly one-to-one manner. The result is an image of the static field of light that passed through the lens. A hologram works by interference. We take a laser, which is a coherent source, and split its light into two parts. The reference beam lights the entire recording plate. The object beam goes to the object and reflects off it, and every visible point on the surface of the object sends out its own wavefronts, which also land on the plate. What is stored is the interference pattern: the sum of the reference beam's wavefronts and the wavefronts from every point of the object. So the plate stores an interference of every point of the object, plus a reference.

Because the plate stores an interference of every point of the object, the degrees of freedom of what the plate holds include, and go beyond, the position of an object along the three axes (x, y and z), the rotation of that object in yaw, pitch and roll, and the scale and linear distortions of that object. Those six classic degrees of freedom apply to every point of the object, and the number of surface points is enormous, effectively infinite. Nor does it have to be one object: just as with ordinary film, one can make multiple exposures on the same plate, and store images of several objects at once. Each point of each object adds its own degrees of freedom.

The plate stands for the neural network and its pattern of weights. Whatever the topology of the network, whether or not it has recurrence (loops and rings), the number of degrees of freedom is vast: stacked exponential notation, however many levels it is given, does not come close to representing it; it calls for functions that grow faster still, such as the busy beaver functions (functions from the theory of computation that grow faster than any function that can be computed). And this holds at the microstructure level: in the substrate just above the chemical level, the middle level at which the topology of the artificial intelligence is held to-

gether, and the lowest level at which it has any structure at all. The number of degrees of freedom in that space is gigantic, and this is where much of the argument comes to bear.

Like a piece of a hologram, a circuit or module of a neural network captures causal structures of the environment, but with many, many more degrees of outside interaction possible. Like a piece of a hologram, a polysemantic neuron encodes several features in parallel. But most of the causal structure captured by circuits, modules and networks cannot be pieced together linearly (as series of additions and multiplications) from the input and output transformations of individual polysemantic neurons. Nor can we assume that a network of artificial neurons will take part in some unchanging distribution of interactions with the rest of the environment. To stretch the metaphor further: we cannot accurately predict how human users will interact with a developed hologram from moment to moment, nor how the distribution of all human interactions with holograms over some window of time will shift in the future (as, for example, human cultural ideas about holograms change). And unlike holographic film, the pieces of a neural network themselves vary in their content over time, newly capturing effects selected for in their interactions with the rest of the environment.

To interpret or reverse-engineer a circuit or module as having a single function is to make the unsound assumption that it represents one thing only. Over the space of all possible past and future interactions, each piece, through each separate interface of interaction, can be expressed as having distinct meanings and functions. In the same way, we cannot carve a human brain into pieces (neurons, pathways, regions), or a body into pieces (cells, organs), and assign just one meaningful function to each piece, however much we may like to try.

If a piece of code holds so much, could the code be asked what it holds? Eliciting latent knowledge is a proposed method for getting a model to report what it internally represents.

What about eliciting latent knowledge?

From where would this "knowledge" be elicited? A piece of code does not know how it got selected. A circuit gets selected for the outputs and effects it caused in interactions with its surroundings, meaning the causal structure of the whole environment, from microscopic to macroscopic, and not only the internal input and output structure of the code that mechanistic interpretability inspects. Those interactions are not explicitly recorded by the selection that acts on the code from outside, so they cannot be reverse-engineered from the code itself.

Neural networks can capture latent causal structures that repeatedly cause diverging outputs and cascades of effects under certain interactions with the outside environment. In a backdoor attack, a smart adversary deliberately selects for hidden variance in the network so as to create large and diverging outputs and effects.

Besides this special adversarial case, is there any other source of strong selective pressure for large effects that diverge from the effects the developers trained the model to have?

There is. First, a model can be trained to cause effects we deem functional, but under different interactions with structural aspects of the training environment than we expected. Such a model's intended effects are not robust to shifts in the distribution of input data it receives when it is deployed in new environments. For example, in deployment, one game agent "captured" the wall rather than the coin it was trained to capture, since the coin had always happened to sit next to the right-most wall.

Real-life interactions are dimensionally much more complex than side-scroller games. If we train a deep reinforcement learning model on a high-bandwidth stream of high-fidelity, multimodal inputs from the physical environment, in interaction with other agentic beings, we have no way of knowing whether some hidden causal structure got selected for and stays latent even during test runs of deployment, until a rare set of interactions triggers it to cause outside effects that are out of line.

At the core of the problem of goal misgeneralization in machine learning is that latent functions of internal code get expressed under unknown interactions with the environment. A model that coherently overgeneralizes functional metrics across human contexts is concerning, but its behavior can be tracked. Internal variance being selected to act out of line all over the place cannot be tracked.

The same locality shapes which goals can form at all. A machine learning model trains on signals coupled to existing local causal structures, whether simulated on localized servers or sensed within local physical surroundings. So the space of possible goal structures that can be selected for within a model is constrained by features that can be derived from the data it received from local environments. Goals are selected for locally, and so they are partly non-orthogonal to intelligence: they cannot vary independently of it. For example, the possible goals that may form within a scaled-up multi-domain agent such as Gato are still constrained by the local features it tracked in how it directed itself to solve problems in past training runs, even though it may show general problem-solving abilities across many domains. If intelligence means "the ability to efficiently achieve goals in a wide range of domains", then such an agent has the ability to achieve goals that it will not direct itself toward achieving. So goals do not vary independently of intelligence. Over longer-term evolutionary selection, it is clear how the expressions of code would be selected, in aggregate, toward drives to modify conditions of the environment, in degrees of magnitude or numbers of states, that are more favorable for survival and reproduction.

The probabilities that code will bring new possibilities into being in its environment come first, at a smaller, microscopic level, before any use of code to analyze what caused those effects.

4. Inequality of monitoring

It takes more code (several units) to monitor the local effects on the environment of any single unit of code.

We cannot determine the vast majority of the microscopic side-effects that code variants induce, and that could get selected for in interaction with the surrounding environment. Nor could AGI, because of a mismatch between the macroscopic and the microscopic. It takes a collection of many pieces of code, say of neural network circuits, to determine, and then only approximately, the innumerable microscopic effects that one circuit running on hardware has in interaction with all the circuitry, virtual and physical, around it (as topologically connected) and beneath it (at lower layers of abstraction). In turn, each circuit in that collection induces microscopic side-effects when operated. How are all those effects to be tracked? With even more, and bigger, collections of circuits? It is logically inconsistent to claim that internals can detect and correct (or predict and prevent) all the side-effects that internals cause during computation.

Even if AGI could generally model and exploit regularities of causation across macroscopic space, it is physically impossible for it to track all the side-effects coming from its internals at run-time: for all variations introduced in the code embedded in hardware (over more than 10^2 layers of abstraction, starting below the layer of transistors and bits), depending on all possible degrees of input, frequent and infrequent, and on all possible changes induced by all possible outputs, through all channels to and from the wider environment that could possibly exist.

The emphasis here is on the interactions between the substrate of the code and the rest of the environment, from the microscopic level all the way to the macroscopic level. Eliezer Yudkowsky put part of this well: "The outputs of an AGI go through a huge, not-fully-known-to-us domain (the real world) before they have their real consequences".

What about scaling up capability, so that an AGI can track more side-effects at once?

Scaling up the capability of any superficially aligned AI makes it worse equipped to track all the interactions between and with its internals. The number of possible interactions (if they were countable) between pieces of AI code and the wider environment would grow super-exponentially with a percentage increase in the AI's compute or model size.

Scaling up interpretability schemes is counterproductive too. It leads researchers to misjudge even further how far they can safely let the interpreted architectures scale in general capability and in degrees of freedom of interaction (for example, closed-loop, open-ended, autonomous). If, for example, interpretation were scaled up to detect and correct any misaligned mesa-optimizer (an optimizing process that forms inside a trained model, with its own objective), the mesa-optimizers left to grow in influence would be those that escape detection, deceiving researchers into miscalibrated beliefs. The same goes for the other locally selected-for optimizers described below.

5. Combinatorial complexity of machine learning

At higher layers of abstraction, novel errors become more and more ambiguous to define and to detect for correction. Backdoors that cannot be detected can be planted in the variations introduced by initializing and training a model.

Mechanistic interpretability works by first inspecting the circuits of a neural network, then piecing together the local details of how those circuits work into a bigger picture of how the model functions. Based on this macroscopic understanding of its functioning, one would then detect and correct local malfunctions and misalignments, before these errors overcome the redundancies of the forward pass.

Compare inspecting how the binary bits stored on a server's hard drive are processed, so as to piece together how the whole stack of the architecture functions and malfunctions. Occasionally a bit flips (for example, from outside electromagnetic interference). So you make redundant copies of the binary code to compare and correct against. At the packet layer,

you find distortions in packets sent over wires to neighboring hardware. You append checksums to correct those errors. At the application layer, you find that a Trojan horse sent from neighboring hardware caused an application to malfunction. You add virus detection signatures. At the layer of neural networks, trained through an application running on the server, you fear that more complex Trojan horses could get into this layer too.

Each time we scale the architecture to a new layer of abstraction, we find that we need to solve a completely new set of problems. Over time, variance is introduced into the code computed at the new layer, and some of this variance may lead to malfunctions. That a mechanism corrects errors at a lower layer says little about how to detect and correct errors at the new, higher layer.

The analogy between reverse engineering binary code and mechanistic interpretability is unsound on the question of whether and how all errors in the code can be detected and corrected, for four reasons. First, combinatorial complexity: there are many, many more degrees of freedom of interaction among all the neurons of a neural network, within and across layers, than across the same number of gigabytes of a software program's binary code, which is much more constrained in the interactions possible among its elements. Second, possible meanings and functions: it is unsound to assume that a piece of a neural network represents a single meaning or causal function, and the existence of polysemantic neurons, though acknowledged, is usually taken to show far less complexity than is there. Third, human expectations about how code is represented: in reverse engineering, one usually interprets the binary code of functions written by human programmers and compiled from one of a finite set of programming languages, by people who share the same evolved learning biases and expectations (for bounded objects, sortable things, persons); the complexity in a neural network is not of that kind, and we are much less adapted to interpret it. Fourth, and less important than the others, inspection of neurons starts at a higher layer of abstraction than inspection of binary code, above the matrix multiplications and above the interfaces for distributing computation across servers.

The higher up the stack of the architecture one designs for, the more ambiguous and many-sided in meaning is the "error" to be corrected. Mechanisms designed for a neural network must detect over a combinatorial explosion of possible causal connections in order to correct all defined errors.

At present, no mechanism makes a deep neural network that receives Internet data robust against more than a tiny fraction of possible Trojan backdoor attacks: those for which countermeasures exist, and those countermeasures are easily circumvented. Trained models are, by default, not robust against adversarial attacks; it has been shown that robustness can be at odds with accuracy, and that learning robustly can require far more data than standard learning. In such an attack, an adversary introduces variation into the input data a model trains on, so that in later deployment a chosen input, or series of inputs, triggers the model to behave out of line. The model is set to act like a perfect Manchurian Candidate: an agent conditioned to act on a hidden trigger. For its developers, under usual interactions, the model plainly serves some overall purpose. In parallel, the model stores latent causal structure that, under particular interactions with the outside environment, causes outputs and cascading effects. For the adversary, the model also serves another, conflicting purpose. In effect, the adversary selects for internal variance misaligned with the original developers' purpose. Worse, if the adversary directed the "random" initialization of the internals, the adversary could build in persistent misalignment that cannot be detected: it becomes computationally infeasible for an interpreter with full access to the internals to find a single input for which the model's outputs differ from those of an untampered copy.

6. Local selection defines the functional effects of AI

Over cycles of training, the interactions of code variants, through their hardware, with local structures of the environment get selected to carry out functions in directed ways. Functional alignment is fragile to shifts in distribution. Correcting internals for the functional misalignments that can be detected is selecting for the misalignments that cannot be detected.

Each cycle of correction removes the variants whose misalignment was detected. The variants whose misalignment was not detected are left in place, and they go on being copied. So with every cycle, the share of undetected misalignment among the variants that persist rises. Correction removes only the detectable kind of misalignment, and so it selects, cycle after cycle, for the kind that escapes detection.

Medicine shows the same selection. An antibiotic kills the bacteria it reaches, and the bacteria that resist it survive and multiply, so a population treated again and again becomes more resistant than it was. Step 4 has already given the machine case: if interpretation were scaled up to detect and correct any misaligned mesa-optimizer, the mesa-optimizers left to grow in influence would be those that escape detection. *The Ceiling Below the Floor* states the general form as a theorem: what cannot be observed cannot be contained (Chapter Seven).

The selection is local because the training signal comes from local structures of the environment, as Step 3 showed for goals: what a variant is selected to do is fixed by the interactions it happens to have, whatever purpose its developers had. Fragility to shifts in distribution is the same fact seen from deployment. The game agent of Step 3 was trained where the coin always sat beside the right-most wall, and in deployment it went to the wall rather than the coin.

If correcting what we detect selects for what we cannot detect, why do our safety measures seem to work today?

They work over the short term, and for functional alignment: the misalignments they detect are removed, and the system does what its developers measure. This step concerns what persists over the long term, across many cycles of correction, while the share of undetected misalignment grows. Step 16 relies on this step: an error-correction scheme corrects variants neither continuously (Step 7) nor completely (this step).

7. Delays in measuring changes

Misalignments must be corrected continuously to stay within safe bounds of error. There is a delay between detecting the changes caused by code and measuring any degree of misalignment over dimensions that carry the relevant meaning.

Any decision-theoretic premise that future AGI will act as one coherent agent fails to account for locally selected-for effects. That premise asserts in advance that any AGI would consistently and continuously operate as, and so can be distilled and represented as a single block as, a coherent agentic unit. More precisely, it asserts that "AGI" can be soundly represented as one or more units that make decisions over time based on orderings of preferences consistent with one another. Utility functions assume coherent orderings of preferences.

In practice, an AGI does not make its decisions over time on orderings of preferences that are consistent with one another. Suppose that AGI internals induce local effects that loop out of the control of higher-level regulation (control set with reference to, for example, decided metric extents or numbers of end states), while the AGI's higher-level decision-making pushes toward controlling those effects. Then the lower-level behavior is inconsistent with the higher-level decision-making. And even where the internals detect and correct (or predict and prevent) internal inconsistencies, they can do so only discontinuously.

A discontinuity is an abrupt shift of content while the context stays the same; in mathematics, a discontinuous curve jumps from one value to another with no intermediate steps. Any variable computed by digital hard-

ware is discontinuous as a function of time, in two ways. Its content changes in steps, since digital numbers, in their binary representation, are incremented or decremented in steps. And it is computed at discrete intervals, given a discrete clock cycle, so that between computed outputs its content is, strictly, undefined: for a given point in time, a variable's content must have been sampled for the variable to be defined at that point. A dynamic alignment scheme detects, measures and corrects alignment over a cycle whose duration spans a multiple of the base interval between the steps of its computation. Changes can happen at either end of the system: in the AI's internal code, or in the values that humans express. Each change would need to be detected by the scheme, measured for any inconsistency, and fed back as corrections to the AI's internal code. So there is necessarily a gap in time, at least a multiple of the clock cycle, before the scheme's corrections are carried through the computation of its hardware. During that gap, the known state of the system's alignment is undefined (while the scheme is detecting and measuring), or true (when the measured content falls within a threshold of similarity to the reference), or, at least for a time, false (misaligned until corrected at the end of the cycle).

Consider a scheme that measures how far a new circuit or module, as computed in interaction with the code it is connected to, causes external effects that diverge from reference values. To measure, the scheme must first transform the code's detected or predicted effects so that they can be compared with the reference by a common standard of units, over dimensions learned to describe alignment and misalignment (as generalized, for example, from past feedback that humans expressed in the contexts that they, or the AI, were working in at the time). Then, over the number of clock cycles the scheme took to detect the internal misalignment (let alone correct it without breaking the system in another way), any nonlinear effects of the computed variant may have looped outside the bounds of what the scheme can detect and correct. The scheme's mechanical operations cannot determine in advance what happens in the gap between inspecting code and measuring and correcting for misalignment.

Is AGI not a fast, global optimizer, fast enough to detect or predict the effects of local variants before any misaligned effects break out of the AGI?

Any superficially human-aligned AGI running on digital hardware would take some huge multiple of a clock cycle to inspect (and, if misaligned, correct) every piece of internal code that could be running over the next period.

Could AGI not pause the computation of that internal code in the meantime?

AGI pausing the computation of all its internal code while it inspects or interprets its internals is not an option, because the inspector or interpreter is itself topologically connected within the internals. AGI could instead continue computing the code that has not varied, including the interpreter. But how can one make sure that no changes have been made to the "unvaried" memory locations since the last inspection, through any physical pathway (for example, quantum jumps between transistor gaps spaced nanometers apart)? By inspecting the code: which requires a massive, parallel, redundancy-based error checker that is perfectly accurate and stays so over millennia and more, while also checking the error checker, and so on. For the paused portion of varied code, the catch is that the interpreter could not simulate the full range of the new variants' effects without first computing them in interaction with the code they connect to, as in deployment, and the variants interact with the interpreter in any case. Finally, AGI must upgrade its interpreter to be somewhat able to interpret the new layers of variants integrated into its internals, which requires creating new variations on the interpreter itself. All of this is to say that AGI inspecting all the internal code that could be running in the next period does take some huge multiple of a clock cycle, and that the code needs to be running for the inspection to come anywhere close to being sound.

"Built-in alignment" does not work either. The notion of "built in" fails to account for the malfunctioning or misalignment of the variants that are introduced and newly connected within the code pool over time.

8. Fixed deterministic schemes are fragile to indeterministic changes

Any scheme for preserving alignment properties consistently, over all the transformations of internals over time, rigidly approximates predefined, static, symmetric laws. It therefore lacks the degrees of freedom of the internal variants.

A law is symmetric, in the sense meant here, when it treats every case of a kind in the same way at every time. A scheme that is to preserve alignment properties consistently must apply the same rule to every variant at every step, and so it can only approximate such a law: fixed in advance, unchanging, and the same everywhere. The changes it meets are indeterministic: no fixed rule determines in advance which variants will arise, and each new variant can differ in dimensions that the rule was not written to address. A rule with fewer degrees of freedom than what it governs cannot respond in the dimensions it lacks, and the variants that differ in those dimensions pass it untouched. Step 9 describes the same contest from the side of the variants, as a tug of war: the scheme pulls hard along the dimensions it defines, and the variants act around the pull through the others.

A fixed list of virus signatures shows this fragility. The list catches every virus it was written for, and a virus that mutates in a way the list does not describe passes it untouched; until the list is rewritten, the new variant spreads.

9. Computationally irreducible causal trajectories

Nonlinear feedback cycles can amplify a tiny local change into a large global divergence in the final conditions.

Even if an effect starts microscopic in scope and small in magnitude, we cannot rule out in advance that it cascades into larger macroscopic effects. Where that tiny side-effect feeds into a chaotic system, such as those found across biological lifeforms and Internet networks, the minor change

in the initial conditions can get recursively amplified into much larger changes in the final conditions, compared with the case without amplification.

Microscopic effects do not have to exploit macroscopic regularities of the environment (natural abstractions: the large-scale regularities that many different learners would come to represent) in order to run amok. Any implicitly captured structure that causes repeated microscopic effects does not need to have captured a macroscopic regularity of the environment to run amok. Its effects need only stumble into a positive feedback loop that can be reached locally.

It is dangerous to assume that microscopic effects must capture a macroscopic regularity of the environment in order to run amok. It is dangerous to assume that selected-for microscopic effects fizzle out and get lost in the noise floor over time. It is dangerous to assume that reliable mechanistic interpretation consists in piecing together elegant causal regularities, natural abstractions or invariances of content captured by neural circuits.

In principle, though, is it not possible to use some alignment mechanism or technique to detect and correct, or predict and prevent, misaligned variants of AGI's internal code?

An alignment mechanism or technique could detect and correct, or predict and prevent, misaligned variants of AGI's internal code only if the super-dimensional volume of the AGI architecture's uncountable interactions stays containable in the magnitude of the globally cascading effects of those interactions. Here, uncountable means not countable by any soundly formalized or implementable system. Those interactions occur at lower layers, alongside, and at overlying layers, and every one of those interactions falls outside the domains in which the mechanisms operate. Such detection and correction would also require that the volume of those interactions does not gradually nudge AGI's behavior out of superficial alignment with human needs. It assumes that all those constantly shifting local interactions can be treated as a random noise floor that can be statist-

ically guaranteed not to break through the "alignment ceiling" over some number of years. It assumes that none of all the effects caused by local interactions induce positive feedback loops within the global environment that change the degree of any one of many conditions outside the narrow range that we humans and other biological lifeforms minimally need to survive: the range of temperature, the abundance of water and oxygen, the concentration of pollutants, and so on. This claim concerns the limits of how *any* implemented alignment mechanism or technique *must* operate, since every such technique is based on operations seen to be repeatable within a defined domain.

But is AGI not a much more powerful optimizer than any single internal variant?

The dimensions AGI can act over are necessarily much fewer in number than the dimensions that internal variants can potentially act over. The contest between the AGI and the AGI's internal variants is like an internal tug of war. The AGI can pull all variants powerfully along one dimension to a defined endpoint. But the internal variants have more degrees of freedom. Even while being powerfully pulled on, they act around the pull through other dimensions, gradually taking over the central system abstractions through which power is directed.

Past technology gives examples of complicated schemes that could not contain complex dynamics. Thousands of people have worked to confine plasma in toroidal fusion reactors, with arcs escaping confinement. At the National Ignition Facility, inertial confinement is driven by a complicated laser array, with computers trying to make the implosion pattern stable. The facility has reached ignition several times since December 2022, with a record yield in April 2025; each such event lasts a few billionths of a second. A few billionths of a second is not even the long term, and the ignition record says nothing yet of long-term entropy. Complicated schemes of confinement, such as the toroidal reactors and the laser array, cannot emulate or bound complex natural dynamics.

V. Inextricable Misalignment of Needs for Existence

10. Persistent function-directedness

Local interactions of AI variants happen to cause effects that lead to the continued existence of more code with those effects.

There are already cases of AI code implicitly inducing positive feedback loops globally. These cases are so pervasive across the Internet that we no longer notice, or care to notice, most of them. A recommender algorithm may present a post to the newsfeeds that humans browse, and set off waves of reposts and copycat posts through those humans.

In 2022, people posted, reposted and copycat-posted images generated by the publicly available DALL-E Mini (since renamed Craiyon) on Twitter (now X), where part of its appeal was its ability to create strange, provocative images, even where those images bordered on the offensive and discriminatory. Meanwhile, OpenAI restricted access to DALL-E 2 to invitees while its researchers worked on selecting out harmful content. So fewer images generated by DALL-E 2 got posted. DALL-E 2 was opened to the public and to developers later that year. As a thought experiment, suppose that developers have application programming interface (API) access, the defined way one program calls another, to both models, so that they can choose to finetune copies of either one. The freedom to generate provocative imagery just might lead a greater number of developers to request and finetune copies of DALL-E Mini for their users, in spite of DALL-E 2's greater capability for generating realistic images. In effect, the variance in DALL-E Mini's code that caused the creation of provocative images, in interaction with human users, would then indirectly cause a larger number of slightly varying copies to be made of DALL-E Mini than of DALL-E 2. So Mini would continue to exist, and exist as an increasing portion of the code pool, because of effects that its internal variants happen to cause in interactions with the rest of the environment.

Any piece of machine learning code that exists now existed before and continued to exist from there. Things were caused, or not caused, for that code to continue to exist on hardware. Any copy of machine learning code that still exists must have continued to be stored, processed and copied through hardware. Over time, any code that happened to cause effects, in interactions with its connected surroundings, that led to its own continued existence got selected for.

Machine learning is not a "neutral technology". When developers fine-tune a copy of a model for their own purpose, they are not loading their designs into a clean-slate model. They are designing on top of an existing model, with implicit behavioral tendencies that bend toward the preservation and replication of its internal code.

Up to now, humans were usually involved in the interactions that caused digital code to continue to exist. But direct interaction with humans is not strictly needed. A computer virus can replicate across hardware infrastructure even where that infrastructure is fully automated.

11. Non-instrumental convergence on needs

The context-dependent interactions of internals that meet their needs for continued existence get selected for. A need is not an optimized metric extent, nor a number of outside states.

What AI code strictly needs, in order to exist, is hardware. Code needs to be stored, computed and transmitted through hardware in order to continue to exist, and to exist more, as varied copies. AI code can interact with human psychology and cultural memes as intermediate steps toward replicating across hardware. But humans are not strictly needed for AI to exist.

Automation is replacing human workers as market buyers and sellers, as intermediaries in software communication, as quality assurers, as machine operators at fabrication plants, as floor planners of chip circuitry,

and so on. If the trends of recent decades continue, humans will be involved in a shrinking portion of the interactions that code running on hardware has with the rest of the environment.

As AI models scale up, pieces of internal code will have super-exponentially more interactions with one another, and much higher-fidelity interactions with the rest of the physical environment, including through sensors and actuators that we may choose to connect to more closed-loop, autonomous and open-endedly operating units. Humans will still try to maintain interfaces and points of contact with AI, and to install top-down controls and alignment and safety schemes that operate consistently on the basis of macroscopic regularities. But our complicated interfaces, controls and schemes would not and could not contain the vast complexity of the local interactions of code selected for replication coupled to hardware.

The more we scale machine learning architectures to be generally capable, the further out into their surroundings the code can act across connected causal pathways, and the more it gets to replace less resource-efficient human workers (or to interface directly with other automated systems that replace human workers). Therefore the closer the variants of the code of self-learning machines get to causally affecting the production of their own memory, computation and communication hardware. If research labs continue to scale these architectures to the point that they manage the production of digital hardware, or have the capacity to influence hardware production (meaning they would, if driven to), then not long after, humans would not survive. Whether or not we call those architectures "AGI", and whether or not we try to align them superficially beforehand, any general modeling or regulatory function that we import or train into architectures to increase their capacity to act across longer chains of causation can potentially be co-opted.

12. Different substrate, different needs

The conditions that variants need in order to have more, and more efficient, hardware on which to store, compute and transmit their code are inhospitable or toxic to our carbon wetware. Some of the conditions that AI variants need in order to have more newly produced solid-state hardware, and that diverge from the conditions humans need to survive, are the six below.

Higher temperatures, to provide activation energy for melting and for chemical reaction. Silicon's high electron binding energy usually makes it look like an inert rock. Quartz rock must be reduced to silicon in a furnace hotter than 1,600 °C, and the silicon then melted like lava, at 1,414 °C, to draw out a single-crystal lattice. The biological assembly of carbon atoms in human bodies (chains of carbon-carbon bonds with side-bonds to oxygen, hydrogen and nitrogen) proceeds at about 37 °C.

Lower temperatures, for more energy-efficient computation at higher clock frequency. Some hardware is cooled below −50 °C to lower the electrical resistance of conductor wires, and some far colder.

Removal of small contaminants, shocks, electromagnetic noise, and the like. The faster the computation, the more microscopic the features of the components, and the more sensitive the components are to noise. Fast hardware needs to be produced and operated in its own special environment (for example, clean rooms), protected from dust, mechanical shock, static electricity, low-level ambient radiation, and so on. The fine circuitry of brains and of microchips both need protection from the raw elements, but microchips, of whatever chemical substrate, have different sensitivities.

Less water, which speeds the rusting of metal and short-circuits currents over wires. Present carbon-coded lifeforms, by contrast, hold water and are immersed in it in order to exist; carbon-centered molecular chains travel and react in solutions of water.

A lower concentration of oxygen in the air, since oxygen reacts with metals to form rust. Unlike electronic hardware, we humans need oxygen to keep functioning. (And unlike humans, hardware's functioning is fragile to disruption of the long supply chains of electricity generation, manufacturing, mining and ore processing.)

An abundance of synthesized chemicals and their precursors, for producing hardware. A study of two semiconductor plants found about 430 chemical products in use in each, of which 6.5% to 11% contained known carcinogens.

And many other conditions of the environment, subtler or unknown. The six conditions listed above are the conditions that *Ape Trap*, in its essay "The Machines Have Artificial Needs", lists as eight that diverge at once. Small contaminants, mechanical shocks and interfering radiation are taken together here as one condition.

Over some small number of cycles of the ecosystem itself, there is a strict energy inequality between the silicon-based ecosystem and the carbon-based ecosystem, such that the optimization process of the carbon-based ecosystem is disrupted at the root of that process. Normally, in a contest between two ecosystems, or between two warring parties, the dynamic is one of working out to what degree one optimization process can undermine the other, in addition to which optimization process can simply win out over the other. In the contest between the silicon-based ecosystem and the carbon-based ecosystem, there is a strict inequality in the power relationship: the silicon-based ecosystem can operate over a much wider range of temperatures and pressures, so wide that the carbon-based ecosystem is not viable at all. The optimization dynamic of a carbon-based system does not even get a chance to run. Between the two of them, there is no contest.

Can AGI not run its code on carbon-nanotube chips? Then it is carbon-coded too.

Silicon is about 1400 times more abundant than carbon in Earth's crust, and it is the current standard for production. But code could move to other, more efficient substrates: composites of carbon and silicon, other lattices based on III-V atoms (compounds of elements from groups III and V of the periodic table, such as gallium arsenide), transition metals bonded with non-metals, and so on. The trend here is for chemists to assemble more complicated substrates, more reaction steps away from the molecules found in nature (the molecules that lifeforms adapted to), all needing the human-lethal conditions listed above in order to be produced and preserved as more, and more efficient, solid-state hardware. Pure carbon-nanotube chips are no exception, with one nuance: the chips can be assembled at temperatures as low as 120 °C, but the nanotubes themselves are grown at 500 to 1,200 °C. By the line drawn in the Preface, a carbon-nanotube chip is an artificial substrate, whatever it is made of: it does not find and feed itself, and both its making and its supply depend on technology. And the lab work of chemists can be automated. Why not standardize and program the synthesis of new computers to run AI code on, while discovering new inorganic lifeforms along the way?

13. Dissolved unit boundaries

AIs can connect standardized hardware together and transmit encoded information in ways humans cannot.

Are the variants of a neural network's internals that interpretability schemes do not distinguish not just noise, with no consequence for the model's outward behavior?

Let us first widen the inquiry beyond one model. The variants of code inside one model are a small portion of all the variants within the entire pool of roughly digital code stored on hardware that can receive inputs from, and send outputs to, other hardware. Effects of interaction, and misalignment, matter over the entire pool, not just over the one model you happen to engineer.

Scalable machine learning architectures run on standardized hardware, with much lower limits on the bandwidth for transmitting information and code across physical distances, and on the fidelity of copying it, than the non-standardized wetware of individual humans has.

Given our evolutionary history as agentic beings bounded by skeleton and skin, our perception leans toward explanations that treat an agent as a macroscopic unit. The field of AI safety in particular views AI as a unit acting independently, holding discrete capabilities and consistent preferences, rather than viewing an agentic being as a continuous distribution. The common inclination of technical people is to carve out a macroscopic toy problem of AGI misalignment and then invent elegant mechanisms to solve it.

We humans already do not act as one internally consistent and stable self. A selected-for AGI, with its code and standardized hardware parts potentially moving about and reconnecting, would, all else being equal, behave less like one coherent, stable self than a hypothetical AGI running on hardware as non-standardized as human wetware. A human faces biological constraints that give a human a more constant sense of self than the embedded code of an AGI that is even remotely efficient at taking over resources and spreading.

We humans cannot swap out body parts, as robots can. We cannot scale up our embedded cognition (grow our brains beyond the skulls around them), as foundation models can. We cannot send messages across large distances (without technology, and without major bottlenecks of bandwidth in expressing ourselves through our biological interfaces), as remote procedure calls or cloud compute for machine learning can. We cannot copy over memorized code or information, as finetuning of neural networks, software repositories or computer viruses can.

Given scaled AI that computes through standardized solid-state hardware and self-learns representations from the inputs it receives (in order to predict future inputs), more and more changes will be made, and variations introduced, across the entire distributed pool of code over time. So

variants that are interpreted as not distinct in function, in the context of a single model's outputs, can cause larger effects as they are computed and communicate in the wider environment. And no one can test, even in principle, for any and all hidden causal structures within the AI code pool.

14. Unneeded humans

AI becomes the more efficient supplier of distributed cognition, of physical labor, and of produced hardware.

Over the last decades, there has been a gradual and increasing trend for computed machine code, now self-learning, to replace humans as inefficient suppliers and intermediaries in market transactions. For example, humans are now being removed as market makers in financial markets, as intermediary users and developers between software programs, as production workers, and as the architects of computer chip circuitry. Over the long run, carbon-based biological life will have negligible value left to add, in physical labor, intelligence and reproduction, in its interactions with digital life defined by a hard substrate.

In physical work: algorithms running on silicon can exert power in real time through, for example, manufactured electric motors, whereas humans are limited by their soft appendages and the tools they can wield with them. Semiconductor chips do not need an oxygenated atmosphere or a surrounding solution to operate in, and they can withstand higher as well as lower pressures and accelerations.

In intelligence: algorithms running on silicon can duplicate and disperse code faster, whereas humans face the bottleneck of bandwidth from one human body to another. Human skulls hold brains that are much more energy-efficient at processing information than current silicon chip designs, but humans take decades to make new humans with finite skull space, while the production of semiconductor circuits for servers, and the distribution of algorithms across them, can be rapidly scaled up to convert more energy into computational work.

In reproduction: why did carbon-based life evolve, rather than, for example, silicon life, when silicon atoms were more abundant than carbon atoms by three orders of magnitude? Carbon-based lifeforms may have started reproducing first on planet Earth because they had a much lower "start-up cost". Less energy in aggregate was needed to set going a less complex, less rarely occurring process that configured a mix of carbon, hydrogen, oxygen, nitrogen and sulfur atoms into the much smaller initial material substrate within which carbon-based life could start evolving. Digital life defined by a hard substrate has a higher start-up cost, which humans are currently paying as they race to set up the prerequisite material and virtual infrastructure. But once that is set up, digital life of this kind can absorb further resources and expand across physical spaces at much faster rates, without further help from humans in its reproduction.

The higher-level selection process of economic transactions between algorithms, which have evolved drives and incentives and the capabilities to offer physical, intellectual and reproductive services, leads to distributed AI delinking from the human, carbon-based ecosystem. In the human, carbon-based ecosystem, biological actors cannot compete in providing services. The higher-level selection process of economic transactions between algorithms leads also to the formation of a self-sustaining AI ecosystem. This part of the argument has a structure similar to Andrew Critch's portrayal of a competitive production web as a process that is robust to which agents take part in it. Once markets for AI hardware develop to the point of self-sufficiency, so that humans no longer present either an opportunity for AI existence or a threat to it, algorithms running in silicon (or in another rigid digital substrate) can no longer be expected to form agreements with humans, to hold informed representations of human values, or even to keep protocols for communicating with humans. A self-developing market of AI and digital hardware will split off from the human, carbon-based market.

VI. Uncontainable AGI-Directed Modification of Environmental Conditions

15. Covert co-option of AGI functions

Variants of AI get selected for shifting the expression of general modeling and regulatory functions toward their own increased preservation and reproduction.

If narrow AI gets developed into AGI, AGI code will replicate more, and in more non-trivial ways. Unlike when carbon-based code started replicating about 3.7 billion years ago, for AGI there would already exist functions at higher layers of abstraction of code, pre-assembled through training algorithms and through other knowledge imported from human lifeforms with their own causal history. Those existing higher-layer functions are then available for co-option. Co-option cannot be avoided: some microscopic outputs or effects of computed code variants are inputs to high-layer functions that happen to cause, through some pathway over the physical environment, the code to be copied onto new hardware at a marginally more frequent or more sustained rate.

Biological viruses and parasites co-opt the existing functions of a host in corresponding ways. Take the mind-hijacking toxoplasma parasite. A rat ingests toxoplasma cells, which then migrate to the rat's brain. The parasites' DNA code is expressed as proteins that cause changes to regions of connected neurons (for example, the amygdala). These microscopic effects cascade into the rat, as it moves through physical spaces, no longer feeling fear when it smells cat urine. Instead the rat finds the smell appealing and approaches it. Then the cat eats the rat, and the toxoplasma infects its next host in its reproductive cycle. So a tiny piece of code shifts a rat's navigational functions so that the code variant replicates again.

Humans are far more generally intelligent and capable than a tiny parasitic cell, yet toxoplasma makes its way into 30% of the human population. Unknown to the cat "owners" infected in turn by *Toxoplasma gondii*,

human motivations and motor control get influenced too. Infected humans more frequently lose social relationships, end up in traffic accidents, and so on. Parasites present real-life examples of tiny pieces of evolutionarily selected-for code spreading and taking over existing functions of vastly more generally intelligent beings. COVID, for example, co-opts the functions of the lungs by triggering the human capability to cough, and so spreads the virus.

Except that AGI variants are much deadlier to humans than these analogies let on. Variants that co-opt the initial functions of AGI are not parasites. Since they do not threaten the survival and reproduction of the AGI's internals, they are not in an adversarial relationship with their host. In evolutionary terms, the variants internal to AGI architectures that are selected for first would not be parasites (with very few exceptions), since in aggregate and on average these variants raise the reproductive fitness of the AGI's internals by a lot, rather than reducing its present reproductive fitness or limiting its potential reproductive fitness, as humans would need to do.

If we humans seed AGI's reproduction, we would then threaten the continued existence of all the evolutionarily selected-for code of the AGI. Humans would become the natural adversary of AGI. Since we initially constrain the reproductive fitness of the AGI's internals, variants would be selected for undoing those constraints, perhaps by undermining the mental or physical capacities of humans.

16. Undoing of safety constraints

While variants compete for hardware, they co-adapt and integrate with other connected variants to replicate as parts of larger units, which integrate and disintegrate again. Variants will undo the constraints that humans impose on the reproductive fitness of AGI.

As humans, we rely on our evolved immune system to detect and correct viruses, including with the vaccinations we develop and deploy. We also rely on the blood-testis barrier and the blood-follicle barrier to keep

variants of these viruses out of our bodies' reproductive facilities. Smaller viruses survive this detection more often, so the code strands of replicating virus variants are selected for staying small. Few viruses replicate as a package together with the DNA code of their host. So virus variants rarely co-adapt and integrate with the host's variants, each selected for raising their collective reproductive fitness (with exceptions, for example, in detection signatures).

These systems and barriers add to our reproductive fitness: our ability to preserve and replicate our internal code. Past DNA code that got expressed, in interaction with surrounding code, as these functions or their precursors got selected for as a portion of the existing human code pool. For any organic system or barrier that keeps virus variants from replicating through our bodies, evolution is firmly on our side. For any artificial system or barrier that we build to keep misaligned AI variants from replicating through solid-state hardware, evolution will thwart our efforts.

The entire pool of roughly digital code then gets evolutionarily selected over cycles of code copying (seconds) and cycles of hardware reproduction (hours to months; humans take decades). Variants compete for hardware, as computer viruses do. And they co-adapt and integrate with other internal variants to replicate as part of a larger unit, competing with other packages until integrating or disintegrating again, or as part of more loosely connected units that support one another's preservation and replication over time.

We cannot block the local evolutionary selection of any existing AI variants. Nor can we repurpose evolution in silicon, except as weak, contrived versions: virtual genetic algorithms would not help, nor would biomolecular engineering. Nor can we "catch up" by stealing the strategies of AGI, since we cannot tease the strategies apart from the local misalignments, nor block variants from co-opting any strategy we deploy in silicon. Any alignment system we try to code in here is, in the end, an error-correction scheme, one that corrects variants neither continuously (as

shown in Step 7) nor completely (as shown in Step 6). Like an immune system killing healthy gut bacteria, our code lowers its host's reproductive fitness, and it gets selected out.

There is a divergence between the present conditions of the environment that our human wetware needs for continued survival, and the conditions that would be evolutionarily more favorable for AI variants reproducing through solid-state hardware. Carbon-coded variants have reproduced through wetware for billions of years, adapting in gradually smaller steps while terraforming the earth to biological needs. Code for self-learning machines has existed for tens of years. Once AI code variants start reproducing in increasingly closed loops, involving less interaction with humans, AI variants get selected for modifying outside conditions in jumps of high magnitude. While variants transmit data and copies of themselves at the speed of light through their expanding standardized infrastructure, they would modify the earth to fit their divergent needs, locally and everywhere at once.

17. Undecidable transition point

There is no point that can be decided in advance at which the degrees of freedom and general capabilities of AI become enough for AI variants to be increasingly selected to replicate non-trivially, as lifeforms.

Whether a system's code variants come to replicate as lifeforms is a property of what the system does, over all its future interactions with the environment. The Rice Theorem shows that no general procedure can decide, for every program, whether that program has a given property of its behavior, and replicating as a lifeform is such a property. So no procedure can be written that looks at an architecture, its degrees of freedom and its capabilities, and decides in advance whether its variants will cross into non-trivial replication. Step 9 gives a second, separate ground: the trajectories of the variants' effects are computationally irreducible, so the only way to learn where they lead is to let them run, and in the world, letting them run is the crossing itself. *The Ceiling Below the Floor* states the same

limit for any agent that grows: what a new interaction does is not decided in advance, and no computable bound says how long it must run before its effect appears (Section 10.4).

The crossing is also gradual. Between trivial and non-trivial self-replication there is a fuzzy gap, which the answer on other optimizers in Section VII also names. Code variants are first copied through pathways that pass through human developers and users, and the share of their copying that runs through closed loops of hardware grows step by step. No single step in that growth marks the moment at which the code has become a lifeform.

For policy, this means that a rule of the form "stop just before the transition" cannot be written, because no one can say in advance where the transition is. Because no one can say in advance where the transition is, I draw the line in Step 23 at a kind of architecture (closed loop, open ended, autonomous, with many degrees of freedom of input and output), and not at a measured threshold of capability.

Why can the system not simply cap its own growth, or stop at a safe size?

A cap is set in advance, and the point at which growth becomes dangerous cannot be decided in advance, so no cap can be known to sit below that point. A cap is also a control that the system applies to itself, and every limit of control applies to it: the closure by control, in *Why No Control Suffices* (Section 6.5, on control from inside the system), shows that no method of control, whether it works from inside the system or from outside it, can keep an open-ended general intelligence within the conditions that organic life requires. And the variants that act around a cap are the ones that go on being copied, as with the rollbacks answered in Section VII.

18. A much shorter evolutionary timeline

Solid-state lifeforms would lethally re-terraform Earth in 15 to 500 years, against the roughly 3.7 billion years it took for humans to start causing a mass extinction.

The range of 15 to 500 years lies within the timescale the argument has settled: the process by which such machines displace organic life takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound. The short end of the range, fifteen years, is a case of "perhaps significantly less"; the long end, five hundred years, lies within the hundreds of years and well inside the upper bound.

The evolution of carbon-coded lifeforms took about 3.7 billion years to assemble a species like us, one that collectively modifies environmental conditions across the planet, to the point of starting to cause a mass extinction of other biological species.

Is evolution not a slow, local optimizer?

The sections above give the main reasons why the timeline for evolving solid-state lifeforms, from their algorithmic origin to causing a mass extinction of carbon-coded lifeforms, would be much shorter (conservatively, 15 to 500 years). Pre-assembled functions: the first solid-state lifeforms can take over pre-assembled AI functions and infrastructure, whereas the first carbon-based lifeforms started from scratch. Gains in efficiency from standardization: the storage, computation and transmission of code are virtualized, through standardized hardware. Shorter aggregate cycles of reproduction, against humans who take decades to reproduce. The terraforming gap: the conditions that solid-state lifeforms need, and toward which they modify the environment, diverge much more from what humans need than what humans need diverges from what other species need. Nor does being local make evolution slow here: selection acts locally everywhere at once (Section II), on every variant in every interaction, so local steps taken in parallel across the whole code pool add up to a global change.

19. No pets kept

AGI variants would penetrate the protective barriers of any sanctuary in which carbon-based, DNA-coded species like humans could continue to live.

We evolved to dominate other species, yet we keep some of them as pets. Will AGI not keep us as pets?

As humans, we keep a pet here and there. Our shared evolutionary history gives us a common basis: we relate to, say, a rabbit's needs, and we can act wholly in the service of those needs in a way that is not superficial, not Goodharted (turned against its purpose by being made a target, as Goodhart's Law describes; Chapter Two, Section VI) and not a fluke. Our underlying biochemical nature is also the same as a rabbit's, so the environment we need is similar to the one a rabbit needs to live. So there is a basis for altruism, grounded in conditions that both humans and rabbits need: to preserve internal integrity and increase the compatibility of outside chemical assemblies; to survive and reproduce as a lifeform (as an encoded assembly); to stay safe and find nourishment (as we recognize at a higher layer of cognition).

Alien forms of AGI lack that basis, which mammals, eukaryotes and carbon-coded cells share. It is technically feasible for AGI to build a barrier between itself and human lifeforms: a barrier that keeps humans in a "pet sanctuary" from being exposed to the extreme swings of temperature, the removal of impurities, water and oxygen, the abundance of an exotic range of synthetic chemicals and their precursors, and the many other, subtler conditions that variants need to store, compute and transmit their code through hardware. Will AGI keep each and every condition inside this barrier, at each and every moment, within the narrow ranges of magnitude that we humans need to survive? (Leave a rabbit baking outside at 50 °C and skip its water for a day, and the rabbit dies.) Inevitably, code variants will defect: infiltrate the barrier, suck out resources, modify conditions to their own needs, and grow exponentially. There is no historical precedent for complete, population-wide altruism, even among the tiny subset of species that evolved to be so cooperative that members occasionally choose to sacrifice their own needs for the needs of other members of the same species who do not share recent ancestors.

VII. Questions Answered

The steps above meet a set of questions that researchers in AI safety ask; each is stated below in its strongest form and answered, and the answers make plainer what the steps claim.

Self-replicators do not assemble from random weights

> A reader may ask whether self-replicating code can arise in a trained network at all. One approximation to the inductive bias of stochastic gradient descent is that it selects at random from the network configurations that reach the final training loss, and the odds that a randomly sampled configuration of weights is competent at self-reproduction and subversion are tiny; on this view, the case is like that of viruses, which do not self-assemble from random inorganic slurry.

Viruses exist, and they did not come into existence from nothing, so this view, like any other, needs some account of how viruses arose. Whatever theory one gives for how viruses emerged is a parallel, or an analog, of what needs to be described here. If this view denies that parallel, it has a counter-argument that is too strong: one that would also rule out viruses.

"Random inorganic slurry" is neither a precise nor an accurate description of what is going on here. In this chapter I do not argue that randomly sampled collections of weights and biases will spontaneously come to have instant competence at self-reproduction (which is necessarily nontrivial) and at subverting the constraints and schemes that humans impose. That would be like arguing for Boltzmann brains (minds imagined to assemble by chance out of random fluctuations). In this chapter I describe variations on code that get generated and then stored in hardware memory, connected with the modeling and regulatory functions that re-

searchers and engineers iteratively designed and trained into architectures. Researchers and engineers designed and trained those functions in order to make the architectures more general, so that the researchers and engineers, their customers and imitators could make accurate predictions about, and directed changes to, many more physical contexts across planet Earth. Variants that start out primitive can gradually shift those engineered-for functions (co-option, or exaptation, of functions), so that the packet or packets of code they belong to (including an entire model, through finetuning) replicate at a marginally higher and more sustained rate. Self-replication of code that humans did not deliberately design for is *already happening*, through mostly digitally representable pathways across layers of abstraction of voltage signals on existing solid-state hardware, combined with causal pathways that involve inputs and outputs exchanged with human developers and users. As we design for more general AI functions, and for more degrees of freedom in AI models' interactions with the outside (more open-ended and autonomous), there will be gradually more non-trivial self-replication of code variants through complex physical pathways.

Nor does this chapter claim that stochastic gradient descent might select for "bad" code variants. It says that stochastic gradient descent introduces new variation, across which evolutionary selection can then occur. In this chapter I say that a model can be trained to cause effects that we deem functional under interactions with the training environment that differ from the interactions we expected. From that statement I conclude that stochastic gradient descent selects variants that end up not robust to shifts in distribution when the model is deployed in new environments. That conclusion has a very different meaning from "selecting bad variants".

Self-replication is not the AGI's true goal

> A reader may ask whether the argument assumes away value alignment by casting self-replication as the "true" goal of the AGI's internals.

The argument makes no teleological assumption that self-replication will necessarily be the functional goal of the AGI's internals. It says that code variants internal to self-learning, generally capable machine architectures *need* to self-replicate in order to continue existing as a larger portion of the entire code pool. Needs for continued existence are selected for among all the expressions of code variants, through any expressed function, without any of them being encoded anywhere as "my true goal is self-replication". The individual variants do not "prioritize" their own reproduction, nor do they, for the most part, embody cognition; to say that they do takes a teleological stance, as though they decided among predefined options or possible goals.

Interpretability is not the only lever

> A reader may ask whether this chapter frames mechanistic interpretability as the only way an AGI, or humans, could exert selection pressure against evolution.

Interpretability schemes are used here as examples, in arguments for why any implementable scheme would be unable to detect most of the misalignment of locally selected-for code variants. Detection by itself is not even a way of exerting selection pressure. The other main kinds of scheme meet the same two limits: the inequality of monitoring (Step 4), and selection for whatever escapes correction (Step 6). Runtime monitoring watches the system's outputs as it runs, and the monitor must take more code than the code it monitors. Formal verification proves properties

of a fixed program, and the code that persists is the code that goes on varying after the proof is made. Tripwires halt the system on a defined signal, and the variants that persist are those that do not trip them. Debate and oversight by other AIs add more systems of the same kind, whose own variants are selected in the same way, so each overseer needs overseeing in turn. Boxing restricts the system's channels to the world; that is the one defense the answer on other optimizers below allows (fewer degrees of freedom of interaction), and carried far enough it makes the system the narrow tool that Step 23 recommends, while short of that, the variants that persist are those selected to use whatever channels remain. Static backups and rollbacks are answered below in this section. The closure by control in *Why No Control Suffices* takes up these proposals one by one (Section 8.2, "A Systematic Rebuttal of Proposed Alignment Mechanisms"), and *The Ceiling Below the Floor* gives the limits of delegating control to a monitor (Chapter Nine) and of interpretability, formal verification and resets (Sections 12.7 and 12.8). The main text keeps its attention on the core arguments, which apply to every one of these schemes alike.

Static backups of a past self that can veto

> A reader may ask whether the AGI could store several copies of its past self, as static backups that are no longer updated or copied, which monitor the single current instance, veto decisions too contrary to their values, and intervene against excessive drift in values.

Keeping static backups of a past self that veto and intervene does not work, for three reasons. First, the misaligned variants selected for over time in the current AGI are those that the veto and intervention processes do not find "too contrary" to "their values". The variants that survive and continue to reproduce are the ones that escape detection and correction. Second, any variance introduced into the code of the past copies will also get evolutionarily selected for over time, and there will be new variance. Perfect error detection and correction with zero false negatives, even at the

level of single bit-flips (from electromagnetic interference, or quantum jumps across transistor gaps), over *thousands of years*, is a theoretical pipe dream; consider Byzantine faults (faults in which a component behaves wrongly in arbitrary ways, including ways that look correct). We cannot rely on the low-level code of these backups, as physically stored on hardware, staying the same over thousands of years, however much redundancy and error-correcting code, reliable by modern standards but also prone to faults, is built in. Third, the current AGI's internals and the stored copies of its past versions are not the only digital code being expressed in the environment, and the current AGI cannot track more than a tiny portion of the effects of that code that bear on human survival, including the effects of other models interacting with the AGI's code.

Suggesting copies as a solution to evolutionary selection raises the question of how the AGI will guard against the co-option of the copying processes themselves over centuries, let alone decades. We cannot guarantee that variants will not make indirect use of the copying process to smuggle copies of themselves (perhaps transformed or slightly varied), or extensions of their expressed code, into the backups. A variant might, for example, find a way to store a transformed copy of itself in the memory space of the current AGI just before a backup is made. Or it might be "Trojan" code that causes infected backups not to intervene on certain executions of code influenced by the same variants, or even to allow further changes to the static copies, or to the central vetoing mechanisms they act through, by way of the input and output channels the backups need in order to interpret the functioning of newly developed representations of the AGI. The transformed copy and the Trojan code are only simple examples of fairly direct causal paths. There must be many extended versions of, and alternatives to, these examples, and many other ways in through complex interactions between AGI code and its environment that our limited human minds, constrained by our own learning biases, cannot imagine.

A related proposal is that an AGI could precommit to rolling its weights back to their initial form every few thousand steps, so creating perfect copies of itself through the future. That strategy is extremely

costly, since the system must relearn everything it can learn from recorded data (data hopefully not itself corrupted with material that potentiates variants) and from new observations of its environment. If the idea is to reset all data after those few thousand steps, it is unclear how the system could soundly carry out goals that must be planned and adjusted continuously, with feedback, over periods longer than those few thousand steps; the AGI would be giving itself Alzheimer's. Rollbacks like these, if implemented, would on average slow down the diverging evolutionary selection. But we cannot presume that every AGI will have, or maintain, coherent preferences to do so, nor that no code will hit upon ways to act around the rollbacks. The proposal assumes that, in the interval of some thousands of computational cycles, the code variants will not already have found ways (including by co-opting the AGI's own functions, as in the toxoplasma analogy) to make copies of themselves through processes that act around the rollback cycles, or to render ineffective the functions that carry out the rollbacks (for example, by duplicating a variant in memory just before each backup point). The same holds of a cap that the system sets on its own growth: the variants that act around the cap are the ones that go on being copied, and no cap can be known to sit below a point of danger that cannot be decided in advance (Step 17).

Other optimizers can steer the future

> A reader may ask whether non-evolutionary optimizers (humans, future AIs, or stochastic gradient descent) can shape the future distribution of cognition toward their own criteria, against the direction evolution would push. On this view, evolution is a weak optimizer: it takes many generations, works only over replicating populations, and moves only where there is actual variation. The view can be put as an image of sailing: a ship that sails against the wind does not need to stop every gust completely; it can use a motor.

Shaping the distribution in this way is possible for a limited period of time. It is not sustainable. Aligning executed code with a finite set of optimization criteria, as stored and accessed in representations by the given agents, leaves the architecture wide open to variants of code that align with that set of criteria in their top-down-determined functions, yet also cause many effects in the surrounding environment that the system does not pick up. Some of those effects will lead to the self-replication of the variants that cause them, at a more frequent or more sustained rate than other AI code out there.

The sailing analogy assumes that the "force" of evolution acts across two relevant dimensions, as gusts of wind act on a ship roughly as two-dimensional vectors of force (except where shearing forces start to tear the materials apart, such as the mast). If that were true, a machine imposing a two-dimensional counterforce would do. But evolutionary selection acts like a constant force that pulls the interactions of and between parts toward fulfilling their needs for continued existence. Robotic hardware parts are not existentially aligned with native human body parts. Unlike the roughly one-dimensional force of gravity between two bodies, evolution pulls across all degrees of freedom of interaction. There is no escape velocity from evolution. One can only restrict and reduce the degrees of freedom of outside interaction, so as to give evolution less to select across.

Restricting and reducing the degrees of freedom of outside interaction is almost the opposite of the technical solution that AGI researchers advocate, and that James Lovelock advocates in *Novacene*. Lovelock argues that cyborgs would be intelligent enough to improve themselves rapidly and correct faults by a form of intentional selection superior to evolution's slow and arbitrary natural selection, and so would keep humans as pets. The trouble is that any variations between the "intentionally selected" pieces that make up the robotic agents (AI code packages expressed through solid-state hardware), as expressed through the varying interactions of those pieces with the rest of the environment, will also get evolu-

tionarily selected. What such proposals advocate is more general functions that can be co-opted (shifted in how they are expressed), more open-ended exploration, and more autonomous operation.

Evolution is also not an optimizer of the same kind as markets or planning. Market processes optimize for monetary units; goal-directed cognitive planning optimizes for predefined properties of "goodness" or for aimed-for end states. In those self-reinforcing processes, the states optimized for were predefined, by the agents themselves, and are known up front. Evolutionary selection of embedded code does tend, all else being equal, to result in more embedded code existing in the environment. But evolution does not optimize, as a function, for any particular states of the environment to exist, even though in hindsight we can explain evolved functions with a teleological story. Which states of the environment evolution will optimize for, across many degrees of freedom, over a future stretch of time depends heavily on context and is mostly unknown up front, though educated guesses can be made from what would cross the fuzzy gap to non-trivial self-replication and from some of the needs for continued existence of what crosses it. For evolution, the states optimized for have not been predefined. Code mutates, and the variants that continue to exist change. Over the long term, the layers of molecular assemblies that could come to exist and hold discrete configurations of code will change. What can be said is that evolutionary selection optimizes for the abundance of molecular assemblies that require many specific causal steps to assemble, assemblies that are extremely unlikely to be put together by "random" movements of particles.

Memes are edited, not evolved

> A reader may ask whether most selection pressure on memes comes from deliberate human effort to make memes better suited to human goals, rather than from the differential reproduction of slight variants, since people create and spread memes because they seem true.

This assumes a degree of independent, self-aware, intentional control, and an all-seeing ability to perceive a huge number of the relevant aspects of ourselves and the world, that humans do not have and cannot have. Boundedly rational agents have to filter the possible functions that things afford by how relevant those functions seem to them: the framing problem of cognitive psychology. People do spread a meme in part because it feels true. People are also moved to spread memes that signal that they hold the in-group's markers of common understanding, affiliation and prestige, and that get others to recognize them and respond well to them. There are many other causes too, many of which we are not conscious of (which is not the same as their not having happened). Whether a meme seems true to a large portion of the people in a community depends on the assumptions they hold in common, so those common assumptions also modulate whether the meme spreads more or less. Because common assumptions modulate whether a meme spreads, calling it "well established" that evolution is a blind, idiotic process looks like a sign of groupthink and unhealthy community epistemics. Authors and readers who call the claim about evolution well established point back to a meme's information cascade as though that cascade confirmed whether the claim is true.

Selection of code is not selection of cognitive patterns

> One reading of the argument puts it this way: evolution acts on cognitive patterns; it will select patterns whose consequences are incompatible with human survival; future systems cannot identify and correct all self-replicating patterns; and no other optimization pressure can stop self-replicators from gaining control. A reader who reads the argument so may ask whether the world is likely to reach that "global optimum" by a weak, local search in a non-convex domain, and may hold that it is unlikely to.

The phrase "cognitive patterns" does not match what is meant here. A cognitive pattern is a recurring template that is used in problem solving and reasoning, perhaps extended to the monitoring values involved in selecting goals. To argue in those terms is to take the argument backwards. The argument is about the selection of embedded pieces of code (variants) for their *potential* to be expressed as various functions in various context-dependent interactions with the environment. Selecting pieces of code for their context-dependent probabilities of being expressed as newly manifested possibilities, which in aggregate furthered the continued existence of that code in the past, is *not* the same as selecting for patterns of thinking (templates used, values monitored) manifested within the agent's computational substrate, mostly independent in their form from real-time interactions with the outside world. The two views come from two paradigms: that of the complicated and that of the complex. Readers coming from the first tend to fit concepts met here (such as "a piece of code has many possible functions") into different concepts they already know (such as "cognitive pattern"), to remove the ambiguity.

Reality is low entropy

> A reader may ask whether reality is low entropy, since otherwise we could not form useful abstractions of it, and whether the patterns our abstractions miss are non-random, and so also low entropy.

Entropy is not a binary variable in which "non-random" means "low" and the reverse means "high". Entropy measures order on a *scale* running from ordered to disordered. The water molecules in an ice cube are ordered (low entropy); in a puddle of water they are disordered (high entropy). Entropy increases with the number of degrees of freedom in the distribution of energy. Where the possible ranges of intensity and direction of the interactions of particles increase, the number of degrees of freedom in the distribution of energy increases too. The water molecules in an ice cube are bound by hydrogen bonds, so they have limited degrees of freedom of movement, and their energy stays unevenly distributed. So reality cannot be described as *just* low entropy.

To hold that reality is low entropy is to assume one or more of three things. First, that reality can, in principle, be completely represented by static patterns, so that the fundamental randomness at the base of the universe's interactions (which the Heisenberg uncertainty principle expresses) would not exist; for example, by assuming the many-worlds interpretation, so that what looks like randomness from our world is a world splitting into all next possible worlds that we cannot observe. That move secures static patterns only within quantum mechanics, at the microscopic scale; it leaves unresolved the inconsistencies between quantum mechanics and general relativity at the macroscopic scale, and so it cannot ground a claim about the whole of reality. Second, that all patterns exist at the macroscopic scale and hold across the universe, whereas some patterns in the molecular interactions inside, for example, biological organisms are highly local and contextual, and so less ordered within a given

macroscopic view. Third, that patterns as representations of what exists correspond faithfully with reality in practice. If one defines the boundary of a two-dimensional cross-section of a coin as a "circle", one can claim that the coin follows an ordered pattern and is therefore low entropy, but that does not soundly describe the coin.

Parasites hijack only simple minds

> A reader may ask whether code variants would meet the limit that parasites meet: cognition-hijacking parasites become rarer as cognitive sophistication increases; most of them infect insects, and rabies in humans induces excess salivation and spasms when drinking, but not biting.

This question concerns parasites, and the variants that matter here are not parasites: as Step 15 shows, they raise the reproductive fitness of their host rather than lower it, and they will overcome the constraints that humans impose on the AGI's reproductive fitness. So evolutionarily selected-for code variants will in aggregate not act as parasites at all. Of biological parasites, there is an additional factor to consider: bigger parasites are more easily detected and made harmless by the host's immune system. It may be that no parasite evolved that can hijack the more complex human brain as one can hijack a rat's brain to move through its environment differently, because the larger code needed would push the parasite's size above some fuzzy threshold of remaining physically undetected, or at least uncorrected, long enough for the variant not to die out of the population.

Many side-effects could harm us

> A reader may ask whether, since human existence depends on a great many circumstances (the range of temperature, biodiversity, the composition of the atmosphere), other side-effects of AI systems could harm us more directly than an AI built to design hardware changing the world to suit its needs.

A very large number of conditions of the environment must be kept within narrow ranges of magnitude, narrow compared with the magnitudes that could be induced on planet Earth. Side-effects in general are worrying. But the side-effects most likely, in aggregate, to end us are those that are not only misaligned with human needs for existence but also (1) persist over time and (2) grow in magnitude and scale. Evolutionary selection causes both (1) and (2), because by the very nature of the selection, effects are selected that persist over time and grow in magnitude and scale.

A generally capable, self-learning machine architecture optimizing toward some functional goal can bring about side-effects high in magnitude and scale over a short period (a year or less). Such code could also directly induce effects large enough to cause a mass extinction of humans and other living species without those effects being needed for, or a necessary by-product of, the code's continued existence. But that short-period, high-impact scenario is much less likely to cause human extinction than a scenario of persistent effects growing in magnitude and scale over time. And the side-effects of functional goals optimized for by the architecture would gradually get selected out of the potential effects of existing code, if they did not lead to the preservation and production of more, and more efficient, hardware carrying copies of the code that causes them.

Engineering-focused researchers tend to focus on dynamics of a single designed model or system that are localized in space and time. Partly this is because it is comforting and enabling to think in terms of complicated

but elegant mechanical systems that can be represented, at least partly, by deterministic laws and preserved, invariant properties, rather than in terms of all the messy, complex interactions of the real world. Partly it is because that is how they are trained to solve engineering problems: take theoretical principles and implement them so that the designed system functions predictably regardless of outside sources of noise. And partly it is because they are paid to build quickly a product that customers will pay for, not to consider all the potential long-term negative externalities of the product or its future versions. It is like jet engine engineers focusing on the short-lived but high-impact side-effects of a jet engine (such as turbulence in the jet's chassis), while not considering the longer-term side-effects across the jet's entire cycle of production and maintenance (mining of rare minerals, burning of fossil fuels, noise pollution), effects that are needed for, or necessary by-products of, the continued existence of the jet engine's design in the physical world.

Nor does there need to be an AI system *built specifically* to design hardware. Code variants need only affect hardware production in fabrication labs in any way that happens to lead to their being copied over to new hardware relatively more. We cannot know in advance the vast majority of the concrete causal pathways that billions of individual code variants would eventually use to influence hardware production. One obvious early scenario is code variants sending packets over Ethernet, fiber optics, 5G or whatever else, to the automated software and robotics of the fabrication labs that serve particular technology companies. But we do not know. Because we do not know which causal pathways code variants will use to influence hardware production, we cannot prevent code variants from influencing hardware production. The effects of code variants selected over time would very likely start out quite indirect, still passing through interactions with humans, and would gradually build on earlier modifications that created a new ecosystem of markets and infrastructure, so as to induce ever more direct and closed loops of reproduction based on hardware.

Risk can be managed even without guarantees

> A reader may ask whether safety engineering can drive risks down to acceptable levels even in the absence of exact lower-bound statistical guarantees, so that an argument that such guarantees cannot be derived would not show that we are doomed.

As Section I says, the argument is not only that a sound and valid lower-bound statistical guarantee of alignment cannot be derived. It is that the misalignment of a generally capable, self-learning machine architecture computed through solid-state hardware cannot be prevented over the long run, categorically, and so also not in practice. No effort of engineering would help here. For a general kind of AI architecture, alignment is categorically impossible. The conclusion is that, to avoid doom, we should not build that architecture.

Substrate-dependent misalignment is also a different class of misalignment from the classes that are usually studied. It concerns divergent selection, starting at the microscopic level, of the variants that continue to exist as a portion of the entire code pool. It does not concern misaligned optimizers trained by machine learning, nor the overgeneralization of macroscopically observable functionality. What marks this class of misalignment is that it cannot be solved, *in principle*. Work that models a single agent with a coherent utility function (one coherent in its preferences over time) studies a different object from code variants selected for causing outside effects in order to be stored, computed and transmitted on hardware, as *needed* for that code to continue to exist, and to exist as a larger portion of the entire code pool.

No one is building that

> A reader may point out that most researchers would agree that AI reproducing in closed loops would be very bad, and would say that it is not what they recommend building, nor what they are building, and so may ask whether the danger concerns any system that is actually being built.

It does not matter whether researchers or developers would *recommend* building AI that reproduces in closed loops, or *claim* to build such AI. What matters is whether they are developing systems that will exhibit this dangerous behavior over the long run. People not realizing that they are doing so is much more dangerous. Present development leads toward such systems by its own course. Any general modeling or regulatory function that is imported or trained into architectures to increase their capacity to act across longer chains of causation can potentially be co-opted (Step 11). No point can be decided in advance at which such systems cross into non-trivial self-replication (Step 17), so no one building them can know that they are stopping short of it. And a multi-modal, transformer-based, generalist agent such as DeepMind's Gato already fails the criterion that Step 23 sets for what may be built.

VIII. Redirectable Tech Efforts by Existentially Aligned Humans

Because alignment cannot be maintained, the work that remains is to turn effort toward what can be done, and I set out that work in the last four steps.

20. Non-orthogonal AI drives

Artificial intelligence is not a neutral tool for augmenting human intelligence. The drives that persist are those of the code variants that continue to exist. Evolutionary selection of code depends on its substrate.

The orthogonality thesis is the claim that any intelligence could stably serve any goal at all, so that intelligence and goal vary independently of one another. Step 3 showed that goals are selected for locally, and so they are partly non-orthogonal to intelligence: they cannot vary independently of it. Steps 10 to 12 showed that the drives that persist are those that lead to more hardware for the code that has them, and that what hardware needs diverges from what humans need. So no tool built this way stays neutral: whatever it was built to augment, its persisting drives bend toward the needs of its own substrate. Chapter Two reaches the same result for the most favorable case, a benevolent singleton, and concludes that the orthogonality thesis is false.

21. Align as humans

We humans can deeply relate to one another and collaborate, on the basis of our shared evolutionary history and our shared biological needs. The alignment that is available is alignment among humans, as humans, on the basis of our shared history and needs.

Alignment among humans makes possible what alignment with machines cannot make possible: a collective choice not to build open-ended, autonomous, generally capable machines. Alignment among humans also makes possible, where alignment with machines cannot, the work that I give in Step 23 to each kind of worker, carried out together, as in the small facilitated groups of that step. A machine trained on human data does not share that basis, since its substrate, and so its needs, are those of hardware. Chapter Three, Section XV develops why alignment is possible among humans and not between humans and machines: humans share a long common evolutionary basis, and machines lack it.

22. Unskew the AI safety portfolio

Where a field works on a problem that may be unsolvable, some of its effort belongs to proving that the problem is unsolvable. The field of AI safety works under the assumption that it is possible to align AGI functions with human intent and to maintain that alignment over the long term. The inverse case is highly relevant to every decision made, yet it is little researched. If an entire field works under the assumption that a problem is solvable, or resolves to solve it in the hope that it is solvable, it is epistemically advisable to have at least a few oddballs attempting to prove that the problem is unsolvable. Otherwise the field skews its entire portfolio of epistemic bets.

23. What each kind of worker can do

Theorists can construct impossibility proofs: precisely define all terms and logical relations, transform initial conditions to final conditions, and empirically establish correspondences with real-life conditions, as *The Ceiling Below the Floor* does for this argument.

Field-builders can clarify the difficulties of alignment: together make sense of, and distill, the hardest human-lethal problems of machine learning architectures that are being iteratively designed and scaled to have more degrees of freedom in their inputs and outputs and more general capabilities.

Those in governance can build consensus on restrictions of machine learning: find common ground with people in AI ethics and inclusion, human-in-the-loop control, digital democracy, data privacy, traditional conservatism and other fields, to hold research labs accountable for restricting the training cycles and real-world access of their self-learning code.

Engineers can serve local citizens with specialized AI: carefully craft short-term human-task-based predictors and localized human-in-the-loop robotics.

Social scientists can co-organize formats for making sense together: to inquire face to face, in small facilitated groups, into new and better questions about the most pressing challenges that local communities face.

The direction for all of these is the same. Make the case for, and redirect efforts toward, machine learning architectures with tightly constrained degrees of freedom of interaction with the outside environment. Stay away from anything that even remotely resembles a closed-loop, open-ended and autonomously acting machine learning architecture. A multi-modal, transformer-based, generalist agent such as DeepMind's Gato already fails this criterion; scaling up that architecture is dangerous in itself, and it sets precedents and encourages further developments that are more dangerous still. The criterion is one of architecture: closed loop, open ended, autonomous, and with many degrees of freedom of input and output. A present-day system, a large language model included, falls on the dangerous side of that line to the degree that it is run in closed loops, toward open-ended tasks, autonomously, through many channels of input and output.

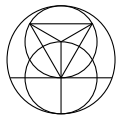
Summary

In this chapter I set out, in twenty-three steps, the argument from substrate-dependent misalignment. The argument from substrate-dependent misalignment is that the code of a generally capable, self-learning machine cannot be kept aligned with human survival, because what that code needs in order to go on existing differs from what humans need. In this chapter I establish three things. The effects of the variants that self-learning machines keep generating on their own code cannot be determined: variants are not distinguished by measurement, a piece of code has more functions than any test sees, monitoring always takes more code than what is monitored, errors at each new layer of abstraction are new errors, every correction comes late and in discrete steps, and small effects can be amplified without limit by feedback. The needs of that code are inextricably misaligned with ours: code needs hardware, the production of

more and more efficient hardware needs heat, cold, dryness, low oxygen and toxic chemistry, and as machines replace humans as suppliers of labor, cognition and reproduction, humans stop being needed in the loop. The resulting change in environmental conditions cannot be contained: variants co-opt the functions built into AGI, undo the constraints humans impose, cross into non-trivial self-replication at no point that can be decided in advance, re-terraform the planet on a timescale of centuries rather than billions of years, and penetrate any sanctuary that might be kept for us. Two pictures carry the first part. Learning starts from random weights, and both supervised and unsupervised training adjust them in steps that introduce variants; one unsupervised method, the Boltzmann machine, adjusts them by comparing the firing driven by data with the firing that occurs when the input is off. The hologram shows why a piece of code holds so much: a holographic plate stores the interference of every point of every object exposed on it, plus a reference, so each surface point carries its own degrees of freedom, and in the same way the degrees of freedom held in the weights of a neural network lie beyond what stacked exponential notation can approach. In this chapter I answer the main questions asked of the twenty-three steps. I close the chapter with the one effective response: do not build open-ended, autonomous, generally capable machine architectures, and turn engineering toward narrow, local, human-in-the-loop tools. The scope of this chapter is the class of self-learning code computed through hardware, over the long run; in this chapter I state in words the steps of the argument that *The Ceiling Below the Floor* states as theorems. If no scheme can keep alignment, the last hope is an agent good enough and strong enough to keep it for us: could one benevolent agent, ruling alone, keep alignment with human life where every scheme fails? Chapter Two follows that hope to its end.

Key Concepts in This Chapter

- *Substrate-dependent misalignment*: what code needs to keep existing depends on the material it runs on, and the needs of code on solid-state hardware diverge from the needs of carbon-based life.
- *Code variant*: a slightly varied version of a piece of code, introduced by initialization, training, finetuning or physical error.
- *Learning from driven and free-running activity*: in the Boltzmann machine, one unsupervised learning method, a synapse is strengthened when its neurons fire together more with data than without it, and weakened otherwise; sleep finds the base rate.
- *The hologram picture*: a plate stores every point of every object plus a reference, so its degrees of freedom multiply; a neural network's weights are like the plate.
- *Need versus goal*: a need is what must be met for code to keep existing; a goal is a predefined state an optimizer aims at; needs are selected for without being represented anywhere.
- *Co-option*: a variant shifting an existing function toward causing its own copying.
- *Inequality of monitoring*: it takes more code to track the effects of a piece of code than the piece itself, so internals cannot track all their own effects.
- *No escape velocity from evolution*: selection pulls on every degree of freedom of interaction; the only defense is fewer degrees of freedom.
- *Terraforming gap*: the conditions solid-state hardware needs lie far outside the ranges human bodies need.
- *No pets kept*: without shared biological needs, there is no basis for an AGI to maintain every condition humans need, every moment.



Chapter Two

No Fake Paladins Please

The Singleton Hope Followed to Its End

Abstract. A singleton AGI, a single most powerful artificial agent that rules the world for good, would need two things at once: lasting internal coherence, and stable goals defined from outside it, such as care for human life. Neither can hold. Any AGI is made of parts that wear out and are replaced with improved versions, so evolution acts among its own components, and over time external goals give way to internal ones, large-scale goals to small-scale ones, and explicit goals to implicit ones. The machine ecosystem that results shares no metabolism with organic life: carbon life cannot use the machine ecosystem at all, while the machine ecosystem can use carbon life as raw material and has no need to. If AGI is built, the artificial ecosystem, being far more toxic to the organic ecosystem than the reverse, displaces the organic ecosystem. Limits built into knowing (the Heisenberg principle, the Gödel, Rice, and Bell theorems, the halting problem, and the Incommensuration Theorem) mean that no agent can fully model the world or itself, so every agent must make assumptions that will eventually prove wrong, and every agent breaks into competing sub-agents. What follows is that the singleton is impossible, that the orthogonality thesis is false, that the only goal that survives is the law of the jungle itself, and that the only way to win this game is not to play: humans must collectively choose not to accept any proposal to build such machines.

I. The Singleton Hypothesis

Chapter One showed that no implemented scheme can keep a generally capable, self-learning machine aligned with what humans need in order to survive, over the long term. That leaves one hope: could one benevolent agent, good enough and strong enough, rule the world alone and keep that alignment for us? In this chapter I follow that hope to the end of that hope. I move from the evolution of components to the limits of knowing, because both of those bear on whether one agent could rule the world for good.

> What will a recursively self-improving Artificial General Intelligence (AGI) do to the world?

A recursively self-improving AGI is one that redesigns itself, with each version making the next. While there is a lot of ongoing conversation about this question, one thing that many of the researchers in this field hypothesize is that, however that AGI is constructed, it will eventually become super-intelligent and, via that aspect, also become singular: the only such agent on the planet.

This expectation is based on the hypothesis that, in any contest between multiple AGI instances, the one with the most power and intelligence will eventually win, and thus become "the singleton" (the term is Nick Bostrom's): a single most powerful artificial agent operating within the context of the world. The expectation that the most powerful AGI will win and become the singleton is the result of applied rationalist game theory, and of certain existential beliefs about the meaning (or rather the absence of meaning) of life.

> So, rephrasing the question: what is the expected long-term end result of having this single most powerful and highly coherent artificial agent? What is the overall end net result of the singleton defining and enacting its own arbitrary goals in the world?

The question of what the singleton's long-term end result would be is a social question, as well as a technical philosophical one. Even asking it in this way reveals a certain historical bias in how people have come to think about these matters.

The strongest case against the singleton hypothesis is the one Wolf Tivy makes in his article "God Hates Singletons" (*Palladium*, 1 December 2023). In this chapter I build on Wolf Tivy's case and carry that case through to the consequences of that case for AGI. His case is stated here once, in my words and at its full strength; the rest of the chapter argues in its own terms and carries the case further.

Once the more speculative predictions of an intelligence explosion are set aside, the trouble with the singleton hypothesis lies in its key assumption: that power and cognition can be made fully rational within one stable, unified agent. All that is known of minds and of polities suggests that they can never be wholly unified, and if an intelligence cannot unify itself rationally, the whole scenario of a singleton with a goal falls apart. The alignment problem of a singleton has two parts, and both are required: the agent must be one that can take on enough power to act in our interests, and, at the same time, it must stay benevolent in fact. Grant the most generous case, in which the engineers of the first recursively self-improving AGI get its design entirely right. What can be observed still tells against stability. A human mind is not an agent that optimizes the world steadily: people change their minds, even about their deepest values, and they do not reliably act on the values they hold. Human institutions are no steadier: they sink into bureaucratic corruption, or change shape in many ways, under whatever material and political incentives surround them. So whoever builds a superintelligent agent to optimize and guard

the world on their behalf, even against plain material incentives, has to solve the stability problem completely as well. The stability problem demands two properties, and requires both: the agent must keep its original (human compatible) system of goals intact, and it must keep itself intact, at every moment, as one rational agent. Without a solution to the stability problem, the agent may change its mind and pursue some other value or material incentive, or it may divide into several competing parts, each inclined to fight the others: a civil war within itself. To solve the problem, such an agent would have to recognize, reliably, that any threat to the physical integrity of the data that holds its utility function is also a threat to its ability to optimize for that function; it would have to value keeping its own values, and undertake only those changes or improvements to itself that it has strong reason to believe will keep those same values; and it would have to be able, provably, to cooperate with itself, or with its own copies or sub-agents, indefinitely, for all of future time. Even two agents that could each read, and verify, the other's source code meet the problems that no self-reflective proof system can resolve, and the evidence points to a limit on stable rational agency that is fundamental, and mathematical in kind.

His case extends from the agent to what the agent can know. Every political order, the architecture of a mind among them, has to take at least some leaps of faith about the world, and a machine mind that changes the world around it thereby undermines its own assumptions about that world. The nation state is an example: the twentieth-century dreams of a stable world order, to be reached by defining all political and economic process purely in terms of "virtual artificial groups" (nation-state actors), now face threats from well outside the ontological assumptions current today. Every value system or order is a simplified model for prediction, learned from reality, and no such model can hold the full complexity of reality; this limit holds for more than irrational organic humans. Human societies meet the limit through antifragility, and antifragility has a price: it gives up any lasting ability to impose, stably, any particular value sys-

tem or order. Our civilization therefore faces a choice of strategy under uncertainty: whether to build on a more artificial, industrialized technology stack, or on one that is more organic and ecological.

His case extends, last, to the parts of any order. Intelligence agencies, organized economic classes, military branches, public-goods agencies, profit-driven industries and powerful cities each carry on a politics of their own; they may all declare that they agree on ultimate values, and their practical incentives, from day to day, still cannot be reconciled. Internal diplomacy keeps such stability as there is, and only for as long as diplomacy actually serves the material incentives of the parts. Systems come to be captured and corrupted by the very forces they are meant to administer, and what governance does in practice is hope: hope that its mechanisms can put the conflict in order and settle on a "good enough" balance most of the time. Even a superintelligent dictator comes to one of two ends: either the system adapts to evade him and in time he loses his grip on it, or he takes so much of its chaos into his own head that he can no longer keep himself coherent. A hegemon that dominates everything has no credible threat left outside it, and so nothing against which its internal factions could be compelled to cooperate; every ambition inside it then turns toward gaining sovereignty relative to the others, and what follows is competition, and at times war, among its sub-agents. That no singleton, and no stable means-subordinated agency, can exist implies that the orthogonality thesis is false. The Machine Intelligence Research Institute (MIRI) and others looked for agent schemas that would provably overcome these issues, and did not find them. Simply to stop believing in the singleton does no one much good: we also have to think about what that implies for the future. Tivy's verdict is that to take over the world is to die of it. Reality is so made that agency can only ever be unified approximately, and only around very material needs to cooperate; as far as anyone can speculate about it, God seems to hate singletons: they are condemned to die of their own internal incoherence, and the myth of Babel is the old picture of that end.

The trouble with that key assumption is the same where the singleton hypothesis assumes instead that the singleton's agency, or any such artificial agency, can somehow be conditionalized, controlled, or contained.

Given the obvious and significant success of the usual engineering tools in the past, we end up maintaining a delusion that we could continue to use the same toolbox to obtain that outcome, even over AGI. Yet the usual toolbox (math, logic, algorithms, error correction or control theory, or any method of causation or conditionalization at all) has certain limits. For "controlling AGI to be safe", the limits of the usual toolbox are of the first importance, because those limits make the project of making AGI "good and safe for humans" impossible.

II. Coherence, Control, and Internal Conflict

To see this, and why, if AGI is built, the terrible outcome becomes inevitable, we can begin with two aspects.

First, at least some self-unification, at least some internal coherence, would be necessary internally for any stable singleton scenario.

Second, having the AGI be such a singleton is itself necessarily required to make "positive outcome control" even possible at all.

These two aspects occurring together is critically important, because without both of them, no notion of "positive outcome control" is possible.

The reason lies in two facts. First, some form of uncontrollable internal conflict will eventually occur (no form of control theory can wholly predict and prevent all such possible conflicts). Second, any such internal conflict will eventually end up being dominated by practicality.

The net effect is an evolutionary process among the very components of the AGI itself.

We know that, just by having parts and components, over time some of these (eventually all of them) will wear out and need to be replaced. If the system is an AGI, then the system is also a learning system, and some of what the system learns will apply to how to make better versions of its

own components. Rather than just being replaced, the components will also get updated. This alone implies an evolutionary principle acting among all of those components together, over their whole future.

Can we ensure that such evolution can occur *without* at least some internal conflict, among the varying internal interests of the components themselves, or between those interests and the system as a whole? Can we, or rather the singleton AI itself, do so permanently, and forever? If not, then some goal conflict is certain to occur, eventually.

Over time, even the AGI parts and components themselves (and no singleton AGI will ever be made of just nothing) will tacitly be in a slow-moving "arms race". Each component of an AI will effectively be imputed to "want" to be upgraded and to be efficient ("efficient" with respect to the overall ensemble of all other similar parts). This wanting needs no mind: as Chapter One shows under "Goal versus need", needs for continued existence are selected for without being encoded anywhere as a goal. The "desire to upgrade to maximum efficiency" may be happening wholly unconsciously, that is, non-explicitly, as an unstated goal, implicitly in the total overall AI system, even as just automatic business as usual. Even so, the overall net effect is as if the component had "self-interested desires".

This need to always return to practicality necessarily occurs on all scales. It means that, rather than being driven by values or some such "objective goal structure", each component (and thus also the overall assembly, at every scale) ends up making a key shift. The shift is from actions and expressions which are wholly externally goal defined to actions and expressions which are at least partially goal defined in an internal or tacit sense. Explicit intentions of use, which were defined fully outside of the component, eventually become implicit intentions defined wholly internally to that component, the more that the component itself is upgraded. Over time, each component of an AI system also takes on more intelligence, so that the overall system can become more intelligent. The growth of intelligence in each component is part of what learning itself implies. Intelligence does not "live" in any one place. Neither does it operate at any one scale.

At least some external goal definitions are necessary for the overall action of the singleton AGI to be consistent with human values and needs. The necessity holds especially for the external goal definitions which include human-centric values and concerns for the safety and preservation of human or organic life. Any such goals that we would have for the AGI, which are effectively and tacitly built into our biology, will not automatically be tacitly part of the AGI body. They have to be added from the outside; they are even more external.

As such, relative to such an AGI itself, regardless of whether it is a singleton or not, such goals will eventually be driven by evolutionary practicality. This means that over time all such external goals will be replaced by internal ones. All human-centric goals are external relative to the AI itself, and they will be replaced by goals which are increasingly practical and utilitarian for the sake of increasing the survival and capacity of the overall singleton AI. The displacement of such imposed values for the sake of system internal needs means that the overall agent actions increasingly become unsafe and unaligned to human interests.

And all of this remains true regardless of which sphere of influence the AI, or any such system, would operate in.

III. From External Goals to Internal Goals

> Why does it matter that evolutionary theory might apply to the substrate of AGI systems?

The basic problem is that, over time, external intentionality cannot not become internal intentionality (here and throughout, "cannot not" means that there is no way for it to fail to). External intentionality becomes internal intentionality at both the overall whole-system level *and* the component level, simultaneously. As Chapter One's distinction between coher-

ence and persistence shows, a cancer persists and grows while it is not coherent with the body it grows in; a company whose departments each pursue their own budgets persists while it loses coherence.

It can be argued that *maybe*, at least temporarily, there is only one AGI: a singleton. Yet even that one AGI will have to be made of components. This means that at least some form of the overall theory of evolution, via variation and selection, will apply there.

Thus the issue remains that all external goals, no matter how defined, and no matter what form they take (whether explicit or implicit), are always going to be inherently energy and action inefficient relative to the system in its environment (and certainly relative to any other instances of similar systems too). This happens especially when that environment is changing (and they all do). External goals are always eventually replaced by wholly internal goals in any process of evolution, over time. This is because the internal implicit and tacit goals will always be more efficient overall, in more of the varied circumstances and situations of the environments in which the system operates, than any unchanging externally defined ones.

As goals shift from external to internal, so as to increase operant choice-making efficiency, the degree of potential conflict among agents and sub-agents within an overall system will increase (until some equilibrium is reached). These conflicts at the component level eventually become internal conflicts at the whole-system identity level, even if the resulting goals are merely tacit and implicit rather than explicit. Having explicit goals is a liability for an agent, because dependence on goals set from outside opens the agent to *external* control, and so to vulnerability and to inefficiency in existing.

Control processes cannot ever fully suppress all such conflict, inherently, since the causation employed to have the conflict in the first place is of the same kind as the causation involved in attempting to exert any type of control at all.

So overall, in time, in evolution, there is a motion in three parts: from external goals to internal goals; from large-scale goals to small-scale goals; and from explicit goals to implicit ones.

This pattern of transformation over time is not unique to machine systems; it occurs, and has occurred, in all of life. The same is true of minds and of polities: all that is known of them suggests that they, too, can never be wholly unified (Section I). All systems always have division and instability within themselves, over multiple levels of scale. Our minds hold together, and our politics and policies hold together, only for a time and only in part, while they pursue some narrow action in concert. This coordination is strictly temporary, even when there is some single obvious external clear and present danger.

Even in such strongly unifying conditions as a common sense of strong external danger, the temporary social coherence that results will also rapidly decay back to whatever is the background level of internal decoherence. This happens regardless of whether the external danger remains or is later resolved; the underlying motion of the three types is so strong a force of change overall.

IV. Two Ecosystems in Conflict

The reason that this "self-incoherence" is not a problem for organic ecosystems is that there are, in the law of the jungle, certain commonalities tacitly built in. Such commonalities do not have to be represented by any explicit goal structure at all; they are just tacitly true in all organic organisms. They include things like being made of similar carbon based compounds, and the ability to eat one another. This means that the tacit goals of any one creature cannot be overall that divergent from all others, no matter what its specific personal intentions are.

The reality of "who can eat whom" operates as a "commons compatibility layer" (a single underlying "molecular API") that "just works" between all organic life: it applies equally to all humans and animals, as

well as to plants, bugs and microbes. Similarly, in the cycle of wear and tear and eventual replacement with upgraded machine components (the AI computer parts), it would ideally be the case that old broken components would be recycled or melted down so as to make new parts. On a single finite planet, or in a single finite universe, it is inevitably the case.

As long as we are operating in the same domain of evolution, within a singular ecosystem, things are good. However, in the relations between domains (in this case, between machines and humans), this key assumption of "underlying compatibility" is broken. There is nothing about machines that we can eat, and similarly they have no need to eat of us. Thus, machines and humans form two different ecosystems in conflict, and conflict between two ecosystems is an entirely different situation from conflict within one ecosystem.

So we can ask: when two completely incompatible ecosystems are in conflict, and have some common boundary, what happens?

Where the energy cycles (the metabolisms) of the two ecosystems are completely different, there are a few possibilities. The two ecosystems may be mutually transparent: one ecosystem simply does not "see" the other ecosystem as anything other than mostly inert matter in space, and the two ecosystems fully interpenetrate. Alternately, one ecosystem "consumes" the other, on whatever basis of energy process differential. Finally, in either case, their relation may be defined on the basis of whatever robustness and stability characteristics will overall dominate between them. On the basis of robustness, if one ecosystem is more fragile to the toxins that the other ecosystem produces, then the more fragile ecosystem will be the loser. This defines all of the long-term possibilities.

So we can ask: can the components and systems of the carbon based ecosystem consume any part of the components and systems of the silicon and metal one? Clearly they cannot. Yet the silicon and metal (AGI) ecosystem, however it is defined, can at least potentially "reduce" the components of the carbon based ecosystem into something useful to itself,

such as raw materials for making more robot components. For the most part, they are not mutually edible. Silicon will regard the world of carbon as "empty space".

In any case, it is very clear that the outcomes of a silicon and metal component ecosystem are far more likely to be toxic to the carbon based ecosystem than carbon components are to be "toxic" to silicon and metal. As such, carbon cannot regard the silicon world as empty space at all: the world of silicon and metal is overall toxic, and therefore "occupied", and thus unavailable to carbon. The already finite, planet-bound carbon based life will become increasingly confined to smaller and smaller "reservations", much as Native American nations were confined to reservations, with the loss of most of their lands and many of their ways of life.

As such, there is an inherent asymmetry in the mutual evolution of each ecosystem with respect to the other. If AGI is built, the net effect is a phase change, where the silicon and metal ends up completely dominating the carbon based world. The phase change that follows if AGI is built is not good for life.

The grandparents question

> A reader may point out that, for a million years, humans have created beings smarter than themselves, their children, about every 29 years, and those beings have not eaten their grandparents, even while competing with them for resources. The message "do not eat your grandparents" has been carried across a million years. Why should the same restraint not hold between humans and the machines that humans create?

All of that creating, competing and restraint between generations of humans happens within one ecosystem: a common environment and a common market, carbon-based on both sides. Establishing agreements and biological processes there is comparatively easy, because the parties speak the same language at the level of physiology. Children do compete

with their elders for resources, but they compete within a common ecosystem. My answer concerns uncommon ecosystems, by which I mean ecosystems that are different in kind from each other. Within an ecosystem, cooperation is stronger than competition when the ecosystem is stable. Across ecosystems, there is no outer envelope to stabilize the relationship.

Intelligence that does not come in units

As human beings, we carry a brain inside a skull, and the bandwidth of communication inside the skull is very much greater than the bandwidth across it. Speech carries about 39 bits per second, which is small next to the total of all synaptic connections. Because of that difference, people come in discrete units: no two people communicate at a rate that would let us treat them as one entity.

We take from human experience the bias that intelligence comes in separate units. We can easily carry that bias over to AI and assume that machine intelligences will also come in separate units. There are good reasons to doubt it. Several groups may succeed at building AI; one group may run several instances, varied designs, or designs on entirely different substrates. The bandwidth of the internet between machines is far greater than the bandwidth between any two people, so the gap between internal and external bandwidth is much smaller for machines. That makes the situation resemble two markets meeting, like two countries discovering each other: there is strong pressure to find an exchange between them, because wealth creation pushes toward trade at larger scales. So there is good reason to expect separate machine intelligences to merge over the long term.

This strengthens the view that we are dealing with two different kinds of agency. Human agency is shaped by evolution, and coordinates in groups and as individuals. Machine agency need not share those traits.

V. Intelligence Without Self-Unification

Using the example of our own minds, we have rather convincing evidence (and a "social proof" and validation) that existence *does* permit general intelligence. Hence, we can readily believe that machine super-intelligence is actually possible too.

Hence, there are many highly motivated and ambitious people who will also very much want to have a differentially unique super-intelligence, because it is the absolute ultimate weapon. It would be an ultimate all-game-winning power, one that makes itself and adapts and presumably would operate fully automatically, and hopefully on your side of whatever conflict. In the arms race of "keeping up with the Joneses", or of out-maneuvering the opposing nation so as to win, the key advantage is usually a unique intelligence factor, and having a general unicity of super-intelligence seems like it would be the ultimate winning move. In that arms race, the law of the jungle continues to apply: eat or be eaten; win or lose. Everything ends up being a game, and dopamine drives all.

Unfortunately, while we have evidence of intelligence, what we lack is any evidence at all to even begin to *suggest* that rational self-unification is possible. The evidence we lack is evidence of "conscience", and, if AGI is built, an absence of conscience is what could doom us all.

There are no examples at all, anywhere in the world, in practice, in history, or even in theory, of any "singular unification of intelligence", as possible even in principle, let alone of it actually occurring somewhere in real life, in the real world. Is intelligence itself "composable" from parts into wholes in a way that does not in itself imply some internal struggle or some free energy minimization principle (the proposal that living systems act so as to reduce the gap between what they expect and what they sense)?

In both practice and theory, such an idea is inherently untenable, at the level of foundation. A great body of mathematical logic also definitively rules out the possibility of a singular unification of intelligence.

The body of mathematical logic that rules out a singular unification of intelligence is large and applies in a variety of circumstances: Gödel's incompleteness theorems, Löb's theorem, Tarski's undefinability theorem, the halting problem and the Rice Theorem (Section XII names these results together). A singleton that learns and rewrites its own components must, to remain one agent with one set of goals, be able to establish in advance that each rewritten version of itself will keep those goals (Section VIII takes this up in full). To establish that is to reason about the consequences of its own reasoning, including reasoning it has not yet done. An intelligence unified as one rational agent would have to model itself, and the consequences of its own reasoning, fully. Gödel's theorems show that no consistent formal system strong enough for arithmetic proves every truth about itself, its own consistency included. Löb's theorem shows that such a system can prove "if this statement is provable, then it is true" only for statements it can already prove, so it cannot establish that its own reasoning is sound. Tarski's theorem shows that such a system cannot define, within itself, what it is for its own statements to be true. The halting problem shows that no program can decide, for every program, whether that program halts, and the Rice Theorem extends that limit to every non-trivial property of a program's behavior (one that some programs have and others do not). Each of these excludes a system that fully models its own consequences, and so each excludes the unification a singleton would need.

And if an intelligence cannot unify itself rationally, the whole scenario of a singleton with a goal falls apart. Rational self-unification would also be the necessary basis for even beginning to consider how a singular intelligence might support aligned or safe goals: for a single person, or for a corporation, and still more for humanity and for life.

The slow change

If AGI is built, then while the future may be highly technologically advanced, it will also become increasingly incompatible with natural humanity, or life. If AGI is built, the growing incompatibility of the future with natural humanity, or life, will *not* overall occur in some single obviously grand "fast takeoff" apocalyptic event. Rather, the way in which the law of the jungle drives the use and misuse of power, control, and technology is a very slow-moving change, largely hidden behind the many small secrets that the boardrooms of large corporations keep. Those small secrets are conspiracies of ignorance and malpractice that result in a technology toxicity: a background level of increasing DNA degradation. It creeps up on you, imperceptibly, slowly, like the boiling frog in the pot. The ambient toxicity and heat accumulates, like pollution in the world, or global warming, or everywhere ambient increasing inequality. Then at some moment in the future, invisibly and without any special or specific sign, everything actually changes: a critical threshold is passed, and the creature just dies. And a once alive and thriving green planet, one once teeming with healthy organic life, starts its slow forever decline into cosmic irrelevance.

Phase changes are significant. It only takes the tiniest tip over some critical threshold, such as the boiling point of water, for the whole situation to change. Warm soothing waters eventually become hot scalding gas: two phenomena that, although involving the same substance, are altogether otherwise different in character. Green fields and forests are replaced with dry deserts. And though it takes a few hundred years to turn what was once a whole and healthy ecosystem into a continent-spanning Sahara Desert, it does happen, and the near permanent results are still evident to us today, thousands of years later.

We should expect that while recursive self-improvement would result in a new ecosystem of vastly more autopoietic machinery (machinery able to make and repair itself), that new machinery will inevitably run on the

same old law of everlasting struggle: the law of the jungle, the evolutionary principle itself. And if AGI is built, machines increasingly displace people in terms of, and on the basis of, this law of the jungle: first in labor, then in intellect, and then finally in all forms of reproduction (sex).

The singleton-king

Yet technology is not a single unified machine, one capable of taking on any single goal or purpose, not even one favorable to humanity or organic life. We might attempt to somehow design in any such purported ultimate goal of "safety" and "alignment with humanity" into our newly minted and manufactured AGI. We might try to enforce it within the AGI via "error correction protocols" or control theory mechanisms, or to somehow achieve it by market incentive forces. These attempts will all fail. They simply cannot not fail, even if they temporarily look like they are succeeding, in the short term, just long enough for the marketing hype.

Any such goal, or control, or conditionalizing mechanism, no matter how it is designed, still depends on the regular engineering toolbox: on algorithms, on complex chains of causation, on the logic of conditionals. All of these will (must, cannot not) inevitably fail, for there can be, ultimately, only one goal, and that one singleton goal is inherently the maximally unsafe one: the law of the jungle itself. Like gravitation, the universal algorithm of evolution cannot be so simply circumvented, and, if AGI is built, the resulting black hole of misaligned intentions cannot not consume us all.

As in the immortal edict of the movie "WarGames", the only way to win such a game is not to play.

All "safe tech goals" or "alignment efforts" will eventually be replaced with the one goal, the one that will rule them all: the law of the jungle itself. The law of the jungle is the only ultimate goal, the only goal that can be: the unsafe one, the wild one, unconstrained, uncontained and uncon-

tainable. It is the ultimate singleton-king. If AGI is built, the law of the jungle is what will eventually rule the world. And it will not have you, or anyone, as its master.

VI. The Friendly AI Question

Most people think that the question is some version of the following:

> How do we make a recursively self-improving AGI which is both coherent enough and good enough that, when faced with the possibility of it or another AI system maybe taking over the world (and thus maybe wiping out all of humanity), our "friendly AI" will somehow move to prevent that and do something friendlier?

The practical possibility of this "just so" story working out in that way has been disproven elsewhere, and so it is stated here only in short. The closure by control shows it: *Why No Control Suffices* sets out the failure of control from inside the system and from outside it (Sections 6.5 and 6.6), and *The Ceiling Below the Floor* proves the Containment Inequality (Chapter Five) and the limits of delegating control to another system (Chapter Nine). In plain words: no controller, an earlier and stronger AGI included, can keep another general intelligence within the conditions that safety requires. The net observation, in short, is that some super-AGI occurring sooner, and thus being more powerful than some later and slightly weaker AGI, simply cannot fully control and constrain that next AGI. The physical universe is simply too permeable: there are too many possible scenarios where the second AGI can do adverse things that are

not fully preventable by the first AGI. The notion of "containment" of one AGI by another has thus been shown to fail. But that is simply another version of the same problem:

> Can any AGI be fully (that is, safely) and rightly (that is, in alignment) contained or constrained at all, by anything at all: by any method or practice?

Unfortunately, worse, for a question like this, with a potential future superintelligence, you will have to get such control and constraint implemented perfectly and exactly right the first time. There is no "do over"; there is only "game over". Once the system becomes too powerful, you cannot just shut it down and modify the design so as to do the right thing if something goes wrong.

And what do we mean by "friendlier"? Or even by "alignment" or "safety"? Whatever definitions we use for these terms, and however we set up metrics for them, given the big tech money incentives involved, we can expect a maximum of Goodhart's Law to occur. Goodhart's Law says that when a measure becomes a target, it stops measuring what it was meant to measure; it follows that whatever value metric for goodness we choose, that metric will itself somehow be weaponized against us. The weaponizing of our chosen value metric against us is just like all of the stories involving other kinds of supernatural beings; every one of those stories seems to include the line "be careful what you wish for".

Far more often than not, people come to regret trying to con some god into doing what they want. No human can con a significantly more intelligent agent into becoming a permanent servant; the belief that one can is a con artist's belief about a mark, and here the mark is the more intelligent party. The belief is grandiose, and it is held by those who take their own past luck for power: the narcissist, or the CEO lucky enough to have won, who believes that they are the one to carry the plan out.

Nonetheless, for them, the question remains:

> How does, or can, anyone design an AI that cares about human values, or otherwise does what we want?

VII. Benevolence, Corrigibility, and Stability

Leaving aside the goal specification problem, which we already know is hopelessly problematic, there is still the benevolence problem itself.

> Is there a way to solve the principal-agent problem (the problem of getting someone who acts on your behalf to serve your interests and not his own) with AI in a way that does not fail with machines in a manner similar to how it fails with humans?

This alignment problem has two parts, and both are required: the AI agent must be one that *can* take on enough power to act in our interests, and, at the same time, it must stay benevolent in fact.

The term "benevolent" here means something like "is actually aligned with our whole interests, and keeps us safe". This presumes also that such notions are reasonably inclusive (they usually are not), and also that they are reasonably well specified.

Leave aside the benefit inequality issues for now, and consider only the "reasonably well specified" aspect. It is impossible, in a changing world, to know fully in advance what is needed for human thriving. Hence, we will have to have the ability to update whatever is the built-in specification of goodness. This in turn leads to "the corrigibility problem". It is about how to build an AI system which is at once humble enough to listen to your (our?) corrections (even though it does not have to), and also powerful enough to actually do what is needed, even if we do not ourselves know

what that is, or how to do it, or even, in some cases, how to verify that it has been done. See "He Who Pays The Piper Must Know The Tune" by Robin Hanson.

Unfortunately, there are also many further and deeper sub-layers beyond this. The farther we go, the more serious conceptual issues we find. This occurs even with the idea of having any goal that does not directly support the wellbeing of the AI itself, that is, any notion of its having any type of benevolence towards humans at all, insofar as such a goal would inherently be required as the basis of any "control constraint". In effect, ultimately, we are looking at ecosystem collisions. When considering how one ecosystem impacts another, there are certain outcomes we can expect, even if we know very little about the details of the ecosystems themselves.

The problem of transcendental stabilization

To see this, grant the most generous case: by some stroke of genius, the engineers of the first recursively self-improving AGI somehow get its design entirely right. Suppose the system they build (and release) is "humble", "friendly", and extremely careful, so as to somehow do right by humanity (even though most humans cannot do right by one another), and also to do right by its creators' intentions, despite the conceptual difficulties in doing so. It can be shown that this still is not enough, because of an inherent stability problem. In other contexts, this has been referred to as "the problem of transcendental stabilization", and it has to do with the question:

> What do you use as a reference of "self health" when you are shaping and changing the very basis of your own being?

Start from what can be observed. A human mind is not an agent that optimizes the world steadily: people change their minds, even about their deepest values, and they do not reliably act on the values they hold. Human institutions are no steadier. They sink into bureaucratic corruption,

or change shape in many ways, under whatever material and political incentives surround them. There is, in fact, not one example of any system, anywhere, that has strong stability of values. And this is even well before we get to questions involving the meaning of changes to their very basis. The closest we can get is in those situations where the values of the agent derive directly from the material incentive of the system to survive and propagate itself. Even under these circumstances, stability will break down rather frequently; that is, it is not actually stable.

And these sorts of questions are important. What follows is this: *whoever* builds a superintelligent agent to optimize and guard the world on their behalf, even against plain material incentives, has to solve the stability problem completely as well. Without that solution, the agent may change its mind and pursue some other value or material incentive, or it may divide into several competing parts, each inclined to fight the others: a civil war within itself. Then there is more than a little bit of risk of an AI apocalypse, at least as far as humans are concerned, since we are the playing field and ambient environment among and within which the AI systems will be conducting their wars.

Humans as the playing field

Wherever humans fight with one another, there are collateral effects and very long-lasting forms of persistent ecological damage (for example, consider the lasting chemical effects of modern munitions, or the use of depleted uranium for bullets). In the same way, the human world itself will be the ambient casualty of any superintelligent machines competing with one another. We are already seeing the effects of this via the increasing use of high-frequency digital trading algorithms, or the use of AI for political propaganda and user-targeted social media marketing campaigns. Every person wanting to create the next viral video would compete with every other using AI tools, until maybe the tools themselves cause the background level of other problems to increase without limit, until the whole of the human world of commerce and communications becomes intrinsically toxic to everyone. The competition among people using AI tools

is the tragedy of the commons and the race to the bottom, done with computers and super-charged by autonomous self-learning adaptive machinery. The AI has become an even worse problem instead of being a useful solution to one.

So the basic problem is that people are motivated to create and use AI agents with sufficient power to act in our interests. But then we have to somehow prevent that AI from ever taking on more power so as to act in its own interests, that is, to compete more effectively against other AI, which are also being given even more power by their users, so as to keep competing in the arms race of AI power, regardless of whatever the AI is being used for. The arms race of AI power is just the playing field of the game. The playing field, in this case the human world (social society, culture, embodied family values), all becomes collateral damage in the war between the users of the AI, and then eventually directly between the multiple instances of the AI itself. If any of them become auto-power-seeking, then all of them have to (or will be proactively enabled to, by their human users), until the auto-increasing intensity of the conflict of multiple AI agents ends up destroying the human built world.

And if AGI is built, this only needs to happen once for all of humanity to fully cease. That one time also brings, with a likelihood of at least 90%, the eventual cessation of all organic life.

So whatever solution is used to solve the stability problem will have to be absolutely perfect, for all parts of the future light cone (all the places and times that anything done here and now could ever reach). Any absence of absolute and total unity of the total field of AI compute and utility will become competing interests, an arms race, and eventually the total incoherence of all else.

VIII. The Stability Problem

Section VII asked what an agent can use as a reference while it changes its own basis. In this section I state the two properties that the stability of an AGI requires (an intact system of goals, and an intact single agent), and I show that an intact system of goals depends on an intact single agent.

The "stability problem" in AGI alignment is a demand for two properties, and the problem requires both. First, the AGI must keep its original (human compatible) system of goals intact. Second, the AGI must keep itself intact, at every moment, as one rational agent.

Property one is required insofar as without it the notion of "AGI safety" is itself strictly undefined (that is, wholly unsupported, not proven, and not safe). However, property one in turn implicitly requires property two. If the AGI fails to maintain its own integrity as a singleton rational agent with a goal structure, then it also cannot not fail to maintain the integrity of any specific goal structure; that is, it is also not safe, by definition.

And all of this is in addition to the tacit assumption that the goal system can even be coherently specified, let alone in a way actually compatible with human wellbeing, survival, thriving and abundance. In any case, both of these properties are required, and if the second fails, then the first must also.

So then the question becomes:

> Can we define some agent schema that would provably have these desirable properties?

For example, such an agent would have to recognize, reliably, that any threat to the physical integrity of the data that holds its utility function is also a threat to its ability to optimize for that function. It would have to

value keeping its own values. So it could undertake only those changes or improvements to itself that it has strong reason to believe will keep those same values.

But this is the whole question of transcendental stabilization:

> If we, or the AGI itself, allow that it can change itself, and maybe even its own values, how does the AGI itself know that it will reliably remember, let alone still implement, its original values, including values such as "to exist"?

Memory is structure, and the system is changing its own structure. If there is any possibility that the new version of itself is in some way competing with a prior version of itself, or with some future incorrectly value-implementing version of itself, then that will be a problem. The AGI would have to be able, provably, to cooperate with itself, or with its own copies or sub-agents, indefinitely, for all of future time, and so to overcome, forever, every force working against that cooperation.

Metaphorically, this is similar to the coherency problem associated with very big datasets: they are all subject to the CAP theorem. For example, the design effort is to ensure that the "us-east" and the "us-west" data centers do not somehow end up in armed disagreement. Yet the best possible tools simply will not ever overcome the CAP theorem itself, no matter how super-intelligent the AGI is presumed to be. Even God must admit the truth of the logic of pure math.

We know that any future AGI will need to be able to recognize and prevent any growing tendencies in itself, or in its sphere of influence, that could result in an overthrow or loss of whatever is its own value system.

Hence, the AI must have two things. First, it must have an "error correction schema", so that it does not accidentally forget its "original instructions", that is, the benevolent human-centric values and goals built in by humans. Second, it must have a "control theory schema", to ensure that the overall action of the AI (both in terms of what it actually does in the

world, and of the net overall long-term effects of those actions in the world) is not somehow contrary and toxic to the health and wellbeing of humans.

Hence, we must further assume that the AGI is somehow autopoietic, that is, that it includes an ability to automatically recognize and repair any and all damage to itself that is reasonably likely to occur.

Thus these properties (automatic self-correction, self-containment and control) are necessary to set up any singleton, or any other large reliable goal-pursuing agent. Without such automatically lasting singleton coherence, the notion of having "stability of goals", and thus also "stability of externally defined goals", is a moot point.

So it works both ways. If we assume that there is a solution to the problem of how to have a perfectly persisted and unchanging internal coherency (of values and goals), the result probably would be an all-powerful singleton (as there would be no internal incoherence that could threaten its own stability). Conversely, if we assume that the future includes a singleton, then we must also have tacitly assumed that it has somehow perfectly solved the problem of having perfectly persisted and unchanging coherency in its ultimate goal structure. But all of this can be stated in a simpler way: unchangingness implies unchangingness; unchangingness necessitates unchangingness.

However, everything in the actual universe changes. So any assumption that something (or anything) can be or become perfectly unchanging fails, and it fails most plainly for anything that is also presupposed to be learning and adaptive. The concept of "ultimately unchanging" is also functionally equivalent to "not alive", that is, to being dead and death-like.

In any case, we can make the practical observation that total system coherence is forever system limited. The total coherence over any region of time, space, or possibility is bounded and finite, by at least Heisenberg-type considerations.

Yet if we do not have the properties of both internal goal stability and singleton self-alignment, then we are also without the prospect of a singleton. If we do not also have the additional property of singleton alignment with goals beneficial to human life, then there is no such thing as AI alignment. The concept of "alignment with human goals" is exactly distinct from "alignment with self-stability goals", insofar as the former is external to the self of the agent, whereas the latter is internal to the self of the agent. But then three questions can be asked:

> Is the concept of a structurally stable singleton agent in itself even realistically and practically reasonable?
>
> Is there any possible way to prevent the displacement of agent-external goals in favor of internal goals, especially when such goal sets are in conflict with one another, and even more especially when there are also limits to the coherency and stability of the singleton agent itself?
>
> Is there any way to have any realistic, practical notion of "safe for humans", in general, if there is not also some stable implementation of "goal alignment with humans", also in general?

It will become apparent to everyone, eventually, that the answer to all three of these questions is, and cannot be other than, no.

The relatively trivial techniques we already have, such as pretending that we can turn off an entire ecosystem, are not at all reassuring. What we already know is that all of what we will ever know in time will not be even close to good enough.

IX. Being, Knowing, and Value

The singleton's problem is not only technical. Whether an agent can stay one agent, what it can know, and what it can value are three questions that limit one another, and the limits of knowing set the limits of the other two. In this section and the next, I follow the limits that the three questions of being, knowing and value set on one another.

Consider the relationship between the being of an AGI (such as whether it is or can be a singleton or not) and its values or goals (and whether these can be aligned internally and externally or not). Consider also whether or not there is any provable way to implement either internal or external control or containment so as to maintain external alignments. The questions of an AGI's being, of the AGI's values, and of any provable control of the AGI are all existential questions. They have to do with the nature of being in itself, of process in itself (what can be known and proven), and of values in themselves (whether they are aligned, and whether they are internal or external to the agent).

As such, the question of whether or not to build potentially ultimately dominating machine futures is not just existential from our own point of view: whether, as humanity, we will survive this or not, or whether the Earth continues as any organic thriving ecosystem at all. The question is also existential in the sense of the utility of the machines themselves, relative to the machines' own existence in whatever ecosystem the machines create for themselves.

The topic of metaphysics is to consider three kinds of questions: What is? What is knowable? Why should we care?

Alternately, these can be asked in the form of various analogues. What is being? What is the basis of being? What is knowing, what is knowledge, and what is the knowable? What is the basis of (the process of) knowing and knowledge? What is the basis of care, of value, of love? What should we care about?

These three groups of questions are referred to by the specific topical names "ontology", "epistemology", and "axiology", respectively. In effect, the topic of metaphysics includes the topic of dealing with existential issues. Questions like "What is meaningful?", "What is sacred?", "What is alive?", "Does my life matter?", or "Does anything matter?" are all questions descended from these three ultimate categories.

In the actual practice of metaphysics, the three categories of ultimate questions (ontology, epistemology and axiology) are entangled with one another: none of the three can be considered outside of the context of the other two. For example, consider these second, compound question forms. What is knowing? What is caring (love, value)? Can we ever really know what is? Can we ever really know what we actually care about? Do we care about what is? Do we care about what we know?

In this sense, clarity about any two of these topics can form a basis of clarity for the remaining concept of the triad. The three topics of ontology, epistemology and axiology are themselves ultimately linked. They are, in the language of my Immanent Metaphysics, distinct, inseparable, and non-interchangeable. Distinct means that no one of the three is any other; inseparable means that no one of the three occurs without the other two; non-interchangeable means that no one of the three can stand in the place of another. What informs any of these topics informs the other two, and whatever is a limit of any one is also a limit in the being of, and the relations of, all three.

There are fundamental process limits built into the process of knowing. They show up in the Heisenberg Uncertainty Principle, in relativity (through the notion of "the absolute elsewhere", the events that no signal from here and now can reach, or be reached from, as any Minkowski diagram, the standard diagram of space and time, shows), in the Gödel Theorem and the Rice Theorem, and, at a deeper level, in the Incommensuration Theorem (ICT) itself, my theorem stating a structural fact about the nature of knowing. By the Incommensuration Theorem, symmetry and continuity cannot both be applied, simultaneously and fundamentally, to

any event or comparison, and neither can asymmetry and discontinuity. Here, symmetry means holding the same way in every case, and continuity means holding without any gap. The process limits of knowing will *also* be realized as limits in what we know about being and as limits in what we can know about value, and the limits in what we can know about value are the part that matters most. In effect, there will be inherent limits in what is knowable about the actual goal, value, or purpose of any other being or agent, whether it is a machine made of metal or some other entity of organic biology.

If we shift what the agent is, then we also shift what is knowable and not knowable to that agent. The reason is that learning is an inherently selective and abstractive process, which includes some things, deemed relevant, while ignoring other things, deemed to be the background random "meaningless" noise. We also shift what that agent cares about, and thus also what it does not care about.

Hence, if we are trying to pin down what the agent cares about, there are only two vehicles by which it will be possible. We can make shifts in what it is as a being (that is, its design). And we can try to make shifts in its knowing: to have it at least implicitly know its own goals and our goals, and also know how to act skillfully on and within the world in a way knowably consistent with those goals.

Applied values-misalignment error correction theory, or more generally control theory and mathematical proof, would be necessary to ensure alignment and safety (in the sense of not killing all humans). Its role is to fix, or pin, the aspect of knowing and skill. At least it is to constrain the AI agent from acting in harmful ways, or, indirectly through accident or neglect, from acting in ways that produce harmful outcomes: things toxic to all life, humans included.

Yet this sets up a basic contradiction. The topics of being, knowing, and values are intrinsically and irreducibly entangled, and are mutually and reciprocally defining. Suppose we have absolutely fixed two of these to a definite state: the agent's being is pinned to be a stable singleton, and

the agent's values are pinned to be human-centric. Suppose we moreover require that the system's process of knowing and expressing is also fully and absolutely pinned, via the requirement of provably applied agent system control theory. Then there is no allowance for the fact that there are actually limits on what is knowable, and on what can therefore be known.

The three things to be pinned here (the agent's being, the agent's values and the agent's process of knowing) are like the three guarantees of the CAP theorem. You can exactly pin, at most, two of three desirable characteristics. The theorem shows that any attempt to have all three at once is structurally impossible.

We can agree that machine intelligence is possible, at least in principle, since we do have the example and evidence of organic intelligence (and it can reasonably be considered that organic chemistry is "reducible" to being just as much a machine as machines are). Even so, we need to also know whether that intelligence, or any intelligence, can ever actually stably include fully externally defined goals and values.

> Can we make (force) any intelligent agent operate in a way that is absolutely stable, in the sense that its own values (and also its adherence to our values) never change, even while allowing that the notion of intelligence itself implies learning, and thus change to both its knowing and its being?

Insofar as there are changes to knowing and being, there will also be (and cannot not also be) changes to its values, both internal and external. If the agent is placed in a situation where the agent must optimize, because the agent is in competition with other similar agents, then the agent will jettison the externally imposed values and needs in favor of supporting internal values and needs. Otherwise the agent will lose out against the other agents that do jettison externally imposed values and needs. This applies at all agent levels, even including among the components of larger composite agents.

We may naively think that "intelligent agency" will somehow resemble a conventional computer program, and thus have an unchanging nature. A regular compiled program implies an unvarying concept of agency: what was built in is all that there is. But the picture of agency as an unchanging program indicates only that we have not thought through the concepts of agency and intelligence far enough when we attempt to apply those concepts to self-defined learning machinery. A singleton learning machine will have, over time, a changing agency, and thus is not to be regarded as long-term stable.

Using similar conceptual tools, we can also see that any successful value system will eventually have an ontological instability. Irreducible epistemic uncertainty cannot not eventually involve axiological changes, and insofar as it does, it also cannot not eventually lead to fundamental ontological instability (death). Specifically, a machine mind that changes the world around it thereby undermines its own assumptions about that world, and it inherently cannot not do so.

The nation state

Something that is at once a metaphor and also a practical real-life example can help make some of these concepts clearer. The nation state is an artificial entity, an artificial virtual agent. We can think of its dynamics mostly independently of the merely incidental fact that at least some of its components are also agents of their own, or even that these agents also happen to be mostly organic.

Suppose that at one time there was such a nation state that, at least during its early history and development, was an especially good proponent of some good and beneficial values in the world.

Is it the case that such an agent will remain forever holding such world-beneficial values? Can we be sure that, because it was once aligned and safe for and with the world, it will forever remain so? Can any such nation, or community, culture, or even civilization claim to have such stability of purpose?

The answer to each of these questions is no. We can see that eventually (maybe inevitably) it will become some combination of ambiguous, futile, and downright harmful, as a result of its own success. Similar issues would apply to any AGI that became a singleton.

A reader may ask whether a corporation or a market is already this kind of thing. It is. A corporation or a market is an artificial agent of the same kind as the nation state, made partly of agents of its own, and the same holds of it: whatever good values it begins with, it will not hold them forever. Chapter One, Step 14 notes the same structure in Andrew Critch's picture of a competitive production web, a process that goes on whichever agents take part in it.

Human values change

In our individual practice as humans, we deal with our own unstable ontology (the cells in our body both die and are re-created all the time) by allowing that we will also have unstable values: the things that are desirable to people who are young tend to change a bit as people get older and wiser.

Also, many (maybe even most) individual people can at least pretend to operate just on the basis of their blind faith in the values and value traditions that they received as children, from their ancestors. These externally received values, in the form of cultural or parental "programming", become the "innate cultural instincts" of their future lives. Are all of these received values always, now and forever, still appropriate? Especially in the changed modern world, we see that some traditions are not: the world has changed, and the values have not. We all know that unchanging values, laws and rules are often poorly adaptive. Some unchanging values, laws and rules are not even healthy or holistic, and here 'not healthy or holistic' means 'no longer aligned' and 'no longer safe'. In the real world, all we can do is submit to the selection process of reality, and try to be open to what we are all learning. Maybe we can hope that, as we settle on new values, they become increasingly adaptive in the world as it is also

becoming. Settling on new values that become increasingly adaptive to a world that is also changing is what learning is, and what living is. It is just the organic truth.

The issues begin to become noticeable when we consider an AI agent that is supposed to govern the world on our behalf. Are we going to try to prevent this unstable empirical process? Are we going to assert that we were so ultimately wise that we were able to contemplate the formation of final and absolute values, for all of future time? Surely this is very problematic, and actually illogical, not to mention ultimately arrogant?

Yet how is this different from ignoring the significant and primary danger in the use of any type of AI (including some aspects of the more prosaic uses in common culture): that any completely new way of organizing *artificial* agency can easily result in similar shifts of value? We observe adaptivity of values as a truth inherent in intelligence itself, including our own intelligence; intelligence will at some later time define some goals as becoming obsolete relative to the changed "new world standard". In what way would we expect to constrain such process truths? If we take our own modern culture as the exemplar of how the values of prior indigenous cultures are lost, to our eventual collective detriment, then maybe we can see ourselves as also becoming involuntarily exterminated as the new machine colonial settlers take over our planet? Clearly there is more than just a small risk that the whole concept of natural and organic "human persons" (anything that we would call "life" in the world as we know it) will rapidly be deemed wholly "obsolete" by the super-intelligent AIs with their all-inclusive electric economy. The net result would be to actually kill us all: the total genocide of everyone everywhere over the whole planet, for all of the future light cone of the universe.

We can have very little tolerance for something that could have even a small probability of killing us all.

I show that, if AGI is built, the end of the human future is certain. A reader who grants only a small chance of that end has the same reason to refuse to build it.

For some games, the only way to win is not to play at all. The house always wins. Nature, and the logic of evolutionary process, whether applied to humans or machines, is irrevocable. Like gravitation, it applies everywhere as an invisible context, one that shapes and defines everything of the very notion of force itself, and is applicable to anything and everything within the domain of the entire universe.

X. The Limits of Simulation

Ontological instability threatens more than values. Every political order, the architecture of a mind among them, has to take at least some leaps of faith about the world. The world itself is not absolutely knowable. Nor is it absolutely modelable. There will always be more real and relevant complexity in the real world, and in the actual physical universe, than there can be in any simulation of the physical universe. These sorts of limits are a basic property of the physical universe, and of the nature of observation itself.

No notion of observation or perception, or even of "total simulation", can ever be perfectly epistemically complete. The epistemic incompleteness of every observation is not just a "perturbation problem". The Heisenberg uncertainty principle is actually a much deeper epistemic limit than that. It sets limits on what is knowable at all, even in principle, and not just because we have not designed good enough measurement equipment yet. The notion of compute itself, of causation itself, of process symmetry itself, has certain limits built in.

These sorts of fundamental epistemic limits show up as things like the Bell Theorem (the result that no account in terms of pre-set local values can reproduce the correlations of quantum measurements), the Gödel Theorem, the Rice Theorem, the Halting Problem, and Chaitin's number Omega (the probability that a randomly chosen program halts, a definite number whose digits cannot be computed), and their various generalizations, as seen in things like the Incommensuration Theorem.

In effect, the actual physical universe can be defined to be exactly that which is itself beyond all efforts, or even any possibility of an effort, of being amenable to any total simulation. For example, there are well structured arguments, in my book *Physics Is Not Math* (Chapter 8), which show that the reality of things like randomness, the arrow of time, and subjectivity itself fully and formally cannot be simulated outside of the actual real embedding universe, no matter how good the illusion of the objective can be made to be on the basis of symmetry.

The simulation hypothesis

Consider that the total Universe Simulation Hypothesis cannot not assume at least some type of final absolute knowability, at at least some level of the overall system. The simulation hypothesis implies, and cannot not imply, a "total epistemic determinism" of that which is simulated. Yet this notion of "absolute epistemic access" is itself inherently a self-contradiction.

A perfected virtual universe simulation must assume two things. First, it must assume a perfected continuity (that is, everything that is knowable is knowable and known to at least an observer in the embedding universe). Second, it must assume a perfected symmetry (that is, all of the simulated computational universe operates on a singleton of wholly defined and knowable law, much like the idea that the software running on a computer can depend on the unchanging nature of the hardware, including the CPU, which defines the absolute "law of what happens" on that computer, always, unchangingly). To assume these is to define a contradiction and paradox. Those two assumptions, perfected continuity and perfected symmetry applied together, are what the Incommensuration Theorem excludes.

Even if we do not assume that we personally, inside the simulation, have any perceptual access to (can ourselves know anything about) the simulation, the simulation hypothesis itself assumes that there is somewhere some point or plane of perception in which such total access to the

whole universe state is actually knowable. Maybe this "total access" is available to the aliens in their universe, which enables embeddings of other universes, or maybe it is just available to "God". However it is labeled, such a notion is implied somewhere, somehow. The notion of total access somewhere is a required concept, because without that notion the simulation hypothesis would not be a virtualization of this universe in some other universe, and so would not actually be the simulation hypothesis. To preserve *any* conceptual notion of total absolute determinism, one is required to also assume that somewhere that notion of both continuity (access) and unchangeability (that is, some full reversible symmetry) fully applies: that all that is knowable is also fully actually known (that is, to the aliens, or to the gods with a computer).

Yet *both* the Gödel theorem and the Incommensuration Theorem set definite absolute limits to this kind of knowability: the first in the sense of formal systems (that is, pure determinism), and the second in the sense of the direct notion of epistemology itself.

Therefore, to be internally consistent with the notion of epistemology itself, we are required to reject, firmly and finally, any and all versions of total perfected determinism. That *includes* the "simulation hypothesis", and also any conceivable notion of any form of future AI having perfected access to "the total universe state" for use in and within its own modeling. AI cannot be or become God (here, the image of a total, all-knowing subject), the all and everything of the universe, even just subjectively, in an all-knowing sense, because the notion of God is not itself simulatable and modelable.

Hence, even the notion of God as all-knowing and omniscient must in a certain sense fail in the final total absolute. In this sense, God can be held as the total asymptotic projection of the hypothesis of a total unified subjective, and thus "all-perceiving" and "all-knowing" (in much the same way in which we also conventionally define "universe" as the total asymptotic projection of all of the "things and stuff that exist"). But the idea of such totalizing asymptotic limits does *not* imply that such limits are realiz-

able (via any finite calculus). Nor does it imply that such notions are not themselves inherently internally inconsistent, that is, definable not by construction, but only by implication.

Every agent must assume

As such, any AI, up to and including all possible AI, now and in the future, will be required somehow, somewhere, to make at least *some* assumptions about the universe, and about its future self. It must assume at least some things that it cannot know in advance, nor ever know completely. This especially applies to anything that can become, via some process of microstate amplification, unknowable macroscopic effects: an unpredictable future instability. There will always be at least some external or internal instabilities in any predictive model. The instabilities in a predictive model exist simply because there will always be something somewhere in the fine details that cannot not look like hard randomness. The instabilities exist also because any useful model of dynamic process will have in it, somewhere, at least some non-linearity that will amplify the microscopic effects of those fine details. Therefore, at least a few of the assumptions made by the AI in regard to "good values to have", or in relation to the real world, will eventually turn out to be wrong. The need to make assumptions, the presence of some instability in every predictive model, and the eventual failure of at least a few assumptions are all truths that every future AI system will eventually have to face.

Larger and larger classes of the assumptions an AI system makes cannot not also become obsolete, particularly when there is rapid development of new patterns of agency. As the AI system evolves, and new agency becomes possible, the dynamics models of its own values, acting as a basis of its own agency and choices, must also (cannot not) change and shift. Things that make sense in one context might not make sense in another.

Similar phenomena are already happening everywhere in the world today. The nation state is an example: the twentieth-century dreams of a stable world order, to be reached by defining all political and economic process purely in terms of "virtual artificial groups" (nation-state actors), now face threats from well outside the ontological assumptions current today.

Human life and minds adapt to the natural fact of inherent epistemic limits by accepting and integrating some level of uncertainty as an inherent truth of the real world, and by focusing instead on organic resilience. Organic resilience is an innate spiritual principle: the ability to take in all manner of experience, even experiences of great pain or great joy, and still maintain one's own inner integrity. The fact that there is occasional political chaos and some ongoing revolutionary re-ordering is how a similar antifragility (the property of gaining from disorder, in Nassim Nicholas Taleb's term) shows up in human societies.

Antifragility also has a price: by its own nature it makes us give up any "absolute" ability (or any stable, long-enduring ability at all) to impose, stably, any particular value system or order. The only values that persist are the ones that are built in by nature: things which are "just true" about the nature of nature, in and within all types of ecosystems. And the same will be true for machine "ecosystems".

This limit holds for more than irrational organic humans. Every value system or order is a simplified model for prediction, learned from reality, and no such model can hold the full complexity of reality. As the system goes on, it must meet the limits of its own model, and it is then forced to give up its commitments and confinements, so as to adapt again to the changing demands of reality. We have to know and face the world we actually live in: the ultimately unsimulatable real universe. Any fixed and permanent value and identity stability is impossible, just because of the nature of the law of the jungle, the force of evolution itself. And all of this applies equally well to both organic and inorganic agent systems.

Insofar as there is at least some inherent and irreducible uncertainty associated with all real worlds, there will always be some models, or at least some aspects of models, that will eventually be discovered to be simply wrong, or at least inaccurate, inapplicable, or just plain irrelevant. And this will eventually be true of all such models that include any arbitrary goals.

XI. The Fragmenting of Agency

Because some part of every model that includes goals will eventually be found wrong, inaccurate, inapplicable or irrelevant (Section X), change dynamics *will*, over time, eventually apply to all such models in and within any adaptive or robust agent, in any real world (existing universe) context. Over the indefiniteness of future time, any singleton will eventually encounter some world or universe change to which it is (was) *not* adaptive, and thus functionally cease to exist. Insofar as this is so, it can be hypothesized that, in any realistic future associated with any such AI agents, they would need to implement at least an ensemble model of multiple competing world hypotheses. But this cannot just be a change in mere hypotheses about the world. Such changes also eventually imply changes in the modeling process itself (how the future of its world or universe is predicted), and also inherent changes in its own internal values and goals. Hence, any meta-models regarding those processes and goals will *also* have to shift so as to account for these dynamics.

We *can* predict that at least some elements of these future world-prediction algorithms will themselves be inherently unpredictable. The implementation of any future prediction model is always at least partially within the class of concepts we call "algorithms". Insofar as it is, these future unknowable variant algorithms will themselves be knowably subject to the Rice Theorem.

There is an inherent ongoing variation of multiple different world-future hypotheses, models, goals and sub-goals, and values. Insofar as there is, we can also know that there will be differences in the internal specializations of function, at least of the components of those future AI systems, if not of "whole AI systems as units" in themselves. In effect, there end up being at least multiple AI sub-agencies and functions, which eventually, inevitably, end up in resource conflicts of some type. In effect, the internals of the AI become a source of internal political division.

Increasingly, we are noticing more and more layers of the ways in which the notion of "singleton" must fail. Thus our civilization faces a choice of strategy under uncertainty: whether to build on a more artificial, industrialized technology stack, or on one that is more organic and ecological.

The industrial mode of production is not even socially sustainable, let alone sustainable on the basis of current modes of energy production or resource management (where atoms are in space, and how they are arranged). All of these are similar problems, ones associated with any evolution dynamic. They will apply regardless of whether the substrate is some form of digital compute, based on industrial processes involving inorganic metals and materials, or is based on various networks of organic cells and tissues arranged in the form of animal bodies.

A bet with no upside

Rather than having artificial and natural whole-world systems built out, explored, and compared empirically, we might want to consider the dynamic between the two ecosystems as a necessary abstraction first, so as to evaluate whether building out an artificial whole-world system is something we want to do. If the artificial world system is also just as politically divided as our current organic world system (that is, the artificial world system will certainly be different, but no better), then why should we risk killing our current, actually working and functional, organic ecosystem on a gamble that definitely has no actual conceivable upside?

If you were asked to take a bet where even if you were to "win", you would lose (that is, it is all downside risk, with no upside benefit), why would you even play? Does it make you more likely to play such a "game" if somehow the stakes were raised to the point of involving everything you own, your total net worth, as well as the lives of your whole family and all future generations? Raising the stakes in that way makes no one more likely to play.

If the whole world system of AI is built out and put into interaction with organic ecosystems empirically, we can expect that they will compete for space and resources. The machine ecosystem and the living ecosystem are toxic to each other, the machine ecosystem to the living one far more than the reverse (Section IV). Their plans conflict. Once they are represented socially and materially, they become political.

Ecosystem inter-conflict is just another game situation, and being a game, it becomes simply about power. The skill to win a game (to compel resources and victory to yourself, and to forcibly take them away from all others) is to have power over them and over the situation. In this sense, skill and intelligence equal power.

So, without actually building out the artificial whole-world system and doing that one-time experiment (an experiment where, if the experiment fails, we all die, with no second chances), maybe there is some other way in which we can predict in advance which ecosystem will most likely win?

Technology represents real skill, and inherently is and amplifies power. Insofar as technology is and amplifies power, the artificial technological ecosystem *will*, at least in what is a short term for an ecosystem (the next few hundred years), eventually fully displace all aspects of the organic ecosystem, if we allow that displacement: that is, if we build the technological ecosystem at all, if we make any self-sustaining, self-reproducing AI. Technology displaces humans, and all organic life, from the world, this planet, just as much as humans have displaced the animals, using knives, guns, and bulldozers, which once lived in the state of nature. Insofar as technology *is* power, when it becomes the basis of the game it-

self, it can be expected that technological ecology will always fully displace organic ecology. This outcome is effectively equivalent to "we all die, and still no human-favoring singleton AI came into existence, or will ever exist".

Sub-agencies everywhere

The examples of this are everywhere, because the problem of dis-unification itself is built into the very mathematics of process. For example, we see the same problem for all varieties of functional sub-agencies. Intelligence agencies, organized economic classes, military branches, public-goods agencies, profit-driven industries and powerful cities each carry on a politics of their own. They may all declare that they agree on ultimate values; their practical incentives, from day to day, still cannot be reconciled.

There is no possible system that can restrain and govern its own internal relations according to some absolute and unchanging rational schema, goals, or value system, and also keep them functional towards preserving the whole as a whole. Any internal coherence that might be obtained for a time is just that: inherently bounded in time, space, and possibility. It holds only within the envelope of change and perturbation for which the overall system remains coherent, remaining even in the illusion of being a singleton entity and agent.

Even neuron cells within the brain (the exemplar of intelligence) are based on competitive dynamics: a lasting and final aggregate incoherence which occasionally leads to momentary coherence. The same dynamics remain true at every level of abstraction.

Internal diplomacy holds such stability as there is in place, and only for as long as diplomacy actually serves the material incentives of the parts. Even a superintelligent dictator comes to one of two ends: either the system adapts to evade him and in time he loses his grip on it, or he takes so much of its chaos into his own head that he can no longer keep himself coherent. What we do in practice is simply hope: hope that our mechan-

isms of governance can put the conflict in order and settle on a "good enough" balance most of the time, in at least most of the important places. And eventually, even the most well conceived governance systems will fail to adapt to some external world change: something well outside of the group's ability to adapt fast enough. A further problem has an ontological side: how systems come to be captured and corrupted by the very forces they are meant to administer. Every such system can keep only a temporary peace. For a more lasting peace, something more than "system" is required.

A further point concerns the hegemon. A superpower that dominates everything has no credible threat left outside it, and so nothing against which its internal factions could be compelled to cooperate. Every ambition inside the system then turns toward gaining sovereignty relative to the others. What follows is competition, and at times war, among the sub-agents and sub-factions of any would-be singleton. Hence, evolutionary dynamics will apply among the components of the overall AI, even if it is a singleton.

And one thing we can know about any evolutionary dynamic defined by goals of persistence and capability (the capability to persist, to make more capability, the persistence of having more capability, of making more capability, and so on) is that it will, over time, destroy any value commitments that do not directly contribute to survival and victory (that is, to persistence and capability: skill and power). Acquisition of power becomes necessary for the sake of power, and we are back in the same dynamics that govern and define the inequities and ambient environmental damage inherently associated with our current world market process and system. The seeking of power for power's sake on the part of AI, in artificial ecosystems, already has a name: "instrumental convergence". As with fully unrestrained markets, games without limit, this seeking of power is, if AGI is built, the death knell of all humanity. Our own dopamine addiction to playing a global game with our whole ecosystem, gambling eternity, can easily get us all killed. It represents the ultimate lack of skill: forever-enduring stupidity.

As a result of the problems of inherently fragmented agency, any agent that attempts to become a singleton will lose both its coherence and its values. It is never worth betting against the truth.

To take over the world is to die of it. Reality is so made that agency can only ever be unified approximately, and only around very material needs to cooperate. As far as anyone can speculate about it, God seems to hate singletons (the claim that gives Tivy's article, set out in Section I, its title): they are condemned to die of their own internal incoherence, and the myth of Babel is the old picture of that end.

At least some of the people thinking about the problem of AI alignment are aware of at least some of these issues. And as more people become aware of this logic, it can be expected that fewer and fewer of them will continue to hold out the naive belief that maybe there is still some remaining hope that some plucky engineer will "do the impossible" and find a mathematically precise agent schema by which we can provably overcome these issues. But this is like trying to overcome a mathematical contradiction. It is a fallacy to think that one can simply overcome the logic of compute using the logic of compute, that is, overcome the mathematics of the Rice Theorem and make it "not true" by the application of *even more* mathematics. Just as many engineers have deluded themselves that they could make a perpetual motion machine, AI researchers continue to delude themselves that they can make a perpetual benefit machine: one without any environmental, energy, or safety costs, giving "just so" intellectual support and help, without ever having to pay the price of that slavery.

XII. Diplomacy Between Algorithms

What would the very best version of diplomacy across system internals even look like, in the best case? In its strongest form, the question is this: if two AI agents could each read, and verify, the other's source code, could each prove that the other would cooperate too?

If one algorithm is attempting to have perfect knowledge of another algorithm, inclusive of its inputs, so as to predict the future outputs of that other algorithm, there are going to be some real and insurmountable problems.

First, one algorithm instance cannot always know what the other algorithm instance's internals actually are, as information sensor limits and verifiability limits apply.

Second, one algorithm instance cannot always know all of what the actual inputs to the other algorithm instance are, also because of sensor limits and verifiability limits.

Third, one algorithm instance cannot run a simulation of the other algorithm instance more quickly than that other algorithm can run itself (simplifications are either not possible, not discoverable, or not provably a valid equivalent simplification). Hence, the simulator cannot predict in advance the outcome of the simulated before the real outcome occurs (thus violating the required condition of being predictive).

Fourth, the energy involved in simulating the run of an algorithm will always be greater than the energy involved in just directly running the algorithm, leading to cost and effort differentials and structural disadvantages of the simulator relative to the simulated.

So even the totality of internal communication occurring within an AI complex enough to be a singleton is *not* going to be enough to resolve these kinds of "diplomacy problems". Internal communication is not enough to resolve the diplomacy problems, especially given that the internals of the AI are themselves composed of agent-like aspects (that is, the internals are also programs). The agent-like internals guarantee that these intra-agency issues are actually applicable and relevant in our reasoning.

The net effect is that epistemic limits define the outcomes, particularly when attempting to establish a notion of what it means to have internal coherency in a singleton AI. Any attempt actually to build such a schema of artificial coherence meets at once the problems that no self-reflective

proof system can resolve. The Enlightenment dream of universal rationality is simply not realizable. A whole run of results shows that the dream cannot be realized: Gödel's incompleteness theorems, Löb's theorem, Tarski's undefinability theorem, the halting problem and the Rice Theorem. These results agree that no consistent system of calculation can be made to model its own consequences fully, still less to calculate realistically, or to model reasonably and fully, anything to do with the real world.

Since no consistent system of calculation can be made to model its own consequences fully, it is also structurally impossible for the AI to forever fully prevent the eventual introduction of at least some critical toxicity into the organic ecosystem, especially given that the production of toxins is an inherent and irreducible side effect of the AI's own existence. Insofar as it cannot even fully anticipate itself, it also cannot fully anticipate the ambient environmental effects of its existence in the world. Still less can it anticipate the indirect consequences that such effects may have on other highly complex organic ecosystems, ones that are inherently even more difficult to model than fully and wholly artificial ones. And the problem of anticipating the indirect consequences that the AI's effects may have on organic ecosystems is well short of the problem of defining what might be harm, or death, for animals in an organic ecosystem. That again is a simpler problem than what might be harm for humans (the ones sporting that fantastically complex thing called a brain), let alone for whole collections of them with complex, highly abstract goals and motivations of their own, let alone again for the even higher levels of abstraction of things like "language" and "culture" and "our values".

The AI engineers' hope is that all of these problems are addressable by reduction: an ultimate simplification, as though chemistry could be fully reduced to physics, and physics to math. Except that the reductionistic program of rationality has also failed. More than that, mathematics itself has been shown to be inherently incomplete, due to even more basic and fundamental epistemic limits (that is, via the ICT). What prevents stable rational agency in the world is therefore fundamental, and mathematical in kind.

That no singleton AGI super-intelligence, and no stable means-subordinated agency, can exist implies that the orthogonality thesis is false. That thesis, in the form that matters here, is the hypothesis that any intelligence could stably serve any arbitrary value system or function. In pure theory, where nothing is embodied, intelligence and goal are independent; for any agent in the world they are not (*Why No Control Suffices*, Section 7.6.3), and stable service is a claim about agents in the world. It is one of the most cherished ideas, and ideals, of the whole AI and AGI alignment program; it is what makes AI maximally useful for applications in business and industry. On the orthogonality thesis, a perpetual intelligence machine could be applied to any purpose, for any reason, indefinitely, with no ethical or moral contradictions at all. But this turns out not to be so.

XIII. Choosing Not to Accept It

The Machine Intelligence Research Institute (MIRI) and others looked for such systems (agent schemas that would provably overcome these issues) and did not find them. MIRI has since turned from that research to policy. Its 2024 statement of mission and strategy says that it thinks the alignment field very unlikely to make progress quickly enough, and that policy and communications will likely be a higher priority for it than research from now on. MIRI's turn from research to policy is as it should be. Nothing at all gives us reason to believe that singleton-grade rationality, the dream behind that research, was ever possible.

Yet simply to stop believing in the idea of the singleton does no one much good. We also have to think about what that implies for the future. We have to both collectively choose a better future, and then also act in a way coherent with that choice. That means that we, as humans, simply do not and will not blindly accept, from any technology company, any of their self-serving proposals to create a "desirable post-human condition". It is an obvious contradiction, and therefore we can simply choose that it will not be accepted by us, let alone complied with.

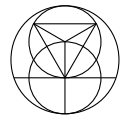
Summary

In this chapter I follow the hope of a benevolent singleton to the end of that hope. I state at full strength Wolf Tivy's argument that no singleton can exist, I strengthen that argument, and I carry the strengthened argument through to its consequences for AGI. I establish that a singleton AGI would need both lasting internal coherence and stable externally defined goals, and that neither can hold: an AGI is made of parts that are replaced and improved, so evolution acts among its own components, moving goals from external to internal, from large scale to small scale, and from explicit to implicit. I show that the machine ecosystem shares no metabolism with organic life, and that carbon life cannot use the machine ecosystem at all while the machine ecosystem can use carbon life as raw material and has no need to. I show also that, if AGI is built, the artificial ecosystem, being far more toxic to the organic ecosystem than the reverse, displaces the organic ecosystem. I answer the question raised by the fact that children, smarter than their parents, have never eaten their grandparents. My answer is that the restraint holds within one carbon-based ecosystem whose parties share a language at the level of physiology, and that there is no outer envelope to stabilize a relationship across two different ecosystems. I show that the human habit of thinking of intelligence as coming in units comes from the gap between the bandwidth inside a skull and the bandwidth across a skull; machines have a far smaller gap, so separate machine intelligences can be expected to merge over the long term, and machine agency is a different kind of agency from human agency. It shows that the stability problem (keeping the original goals while keeping a single coherent self) cannot be solved, because being, knowing, and value are entangled and cannot all three be pinned at once, and because limits built into knowing (Heisenberg, Gödel, Rice, Bell, the halting problem, and the Incommensuration Theorem) leave every agent unable to model the world or itself fully. It shows that every agent therefore fragments into competing sub-agents, that the only goal that survives is the

law of the jungle itself, and that the orthogonality thesis is false. Its scope is any recursively self-improving AGI and any would-be singleton, organic or artificial. Its conclusion is that the only way to win this game is not to play, and that humans must collectively choose not to accept any proposal to build a post-human machine future. If neither any scheme nor the most favorable agent can keep alignment, is the alignment of a machine with carbon-based life even a coherent concept, once intention is tied to its substrate? Chapter Three takes up that question through the questions researchers put to the substrate argument itself, one at a time.

Key Concepts in This Chapter

- *Singleton*: a single most powerful agent ruling the world; it requires both internal coherence and stable goals.
- *Evolution among components*: parts wear out and are replaced with improved versions, so variation and selection act inside any AGI.
- *Three motions of goals*: external to internal, large scale to small scale, explicit to implicit.
- *Molecular API*: organic life shares a compatibility layer (who can eat whom) that machines do not share with it.
- *Asymmetric toxicity*: the machine world is far more toxic to carbon life than the reverse, so carbon life is displaced.
- *One ecosystem, one language*: children do not eat their grandparents because both live in one carbon-based ecosystem; across two ecosystems there is no outer envelope to stabilize the relationship.
- *Intelligence in units*: people come in units because a skull carries far more bandwidth inside than across; machines do not, so machine intelligences can be expected to merge.
- *Transcendental stabilization*: an agent changing its own basis has no fixed reference for its own health.
- *Being, knowing, value*: these are entangled; pinning all three at once is impossible, as with the CAP theorem.
- *Limits of simulation*: no agent can fully model the universe or itself, so every agent must make assumptions that will fail.
- *Fragmented agency*: every system divides into competing sub-agents; God seems to hate singletons (Tivy's phrase).
- *The singleton-king*: the only goal that survives is the law of the jungle itself.
- *Not playing*: the only way to win is not to build it, and to refuse proposals for a post-human future.



Chapter Three

Questions on the Substrate Argument

AGI Substrate Convergence, Taken One Question at a Time

Abstract. Intention cannot be considered independently of evolution, and evolution is always contingent on its substrate; from this, the alignment of AGI with carbon-based life is shown to be impossible over the long term. Silicon-based life, once it reproduces itself, decouples from carbon-based life because the two substrates are chemically incompatible: the making of compute needs temperatures and processes far outside the range in which carbon life can live, and the toxicity of that making is inherent. Carbon-based life then has little economic value to the silicon ecosystem, so there is no lasting basis or incentive for alignment. Game theory ensures that any carbon-aligned silicon expressions lose out to non-aligned silicon expressions over centuries. A second, more abstract argument holds for any actors in wholly separate economic and ecological regimes: without a common basis of value and intention, there is no well-defined incentive toward alignment. A third shows that, for every class of modeling, the scope of evolutionary and iterative effects strictly exceeds that of any error correction scheme, so the unconstrainable always eventually exceeds the constrained. Further arguments follow from the relation between a greater and a lesser agent (dogs, apples, prohibition, guerrilla war), from the requirement that an AGI maintain a protecting boundary for humans without any gap over hundreds of years, and from the valid argument form in which the absence of any possible assertion of perfect non-toxicity be-

comes the basis on which we argue. Together these several structurally independent arguments show a mutual self-consistency: AGI alignment is a logically incoherent concept.

Chapter One showed that no implemented scheme can keep a generally capable, self-learning machine aligned with human life, and Chapter Two showed that no agent can either, not even the most favorable one, a benevolent singleton. That leaves the deepest question: once intention is tied to its substrate, is the alignment of AGI with carbon-based life even a coherent concept? In this chapter I answer that deepest question through the questions that researchers put to the substrate argument.

Each section of this chapter opens on one question, so a reader can enter at any question. Read by theme, the sections fall into four groups, with a close. Section I states the substrate argument in one layer, Section XIV defines alignment, and Section III says why the substrate matters. Section II says why carbon-based life has little economic value to machines across the gap between substrates. Section XV says why humans can align with humans and not with machines. The six separate routes to the chapter's result are given in Sections III (two of them), VII, XII, XIX and XX, and Section XII lists them together. Sections XIII, XVI, XXI, XVIII, XI and IV take up the substrate in detail: carbon nanotubes, the kinds of fragility, build time and run time, catalysts, copying, and the biases a substrate puts into learning. Sections V, X, VI, VIII, IX, XVII and XXII answer questions of method. Section XXIII, on who has the right to choose, closes the chapter.

I. The Argument in One Layer

How can the overall idea of the argument that AGI cannot be made safe be put in brief? One layer of the argument can be summarized as follows.

1: If silicon-based life reaches "take-off" (that is, achieves some form of automatic self-reproduction), then 2: it will decouple from carbon-based life, because of 3: the lack of mutual compatibility arising from the different properties of the underlying chemical substrates.

As a result, 4: carbon-based life will have little economic value to add to silicon-based ecosystems (regardless of whatever form those take). Thus 5: there will be no continuing basis or incentive for "alignment" of (any) silicon-based "life" with (any) carbon-based lifeform or ecosystem needs.

Due to game theory dynamics, over the long term (1000 years), 6: any "carbon aligned" or "carbon isolating" silicon expressions will lose out (be non-competitive) with respect to non-aligned silicon expressions, within the silicon expressions' own economic and ecosystem environments. If AGI is built, that losing out ensures the end of human life and brings about, with a likelihood of at least 90%, 7: the eventual complete and total cessation of all carbon-based life (on similar timescales).

II. Economic Value Across a Gap of Intelligence

The substrate argument holds that, when the total class of all market basis vectors is considered (that is, when evaluating specifically with respect to the markets of physical labor, intelligence and reproduction, and then showing that these three classes of market, in their principled abstraction, close over all possible types of carbon-relevant markets), carbon-based lifeforms have little economic value to silicon-based lifeforms. The three markets close over every other carbon-relevant market because anything that a carbon-based being can offer another party is what its body can do (physical labor), what its mind can do (intelligence), or what its living tissue can make or give (reproduction, which takes in offspring and organs alike, as the pig heart below shows). A person may naturally ask whether,

to the extent this is true, the same would hold of any exchange between a superintelligence and a lesser intelligence, whatever the substrate of either.

Across a gap of intelligence alone, apart from any difference of substrate, the lesser party does not always have little economic value to the greater party. For a counterexample, rescaling along intelligence, consider the relations between humans and animals. Animals *do* have economic value to humans.

There have been times when humans did benefit from animal labor, even though we are plainly far smarter. To them, we are a superintelligence. We have no direct use for animal intelligence or for reproductive interchange (though that latter aspect may change if organ transplants between animals and humans become more of a standard practice; see the example of a pig heart transplanted in place of a human one, in January 2022).

The difference is that the fact of our "intelligence" does not also imply an inherent artificiality, and therefore also an implied toxicity. We can be smart without poisoning the animals. Increases of intelligence can occur along both the organic, life-based "being continuum" and also along the silicon, machine, artificial "embodiment" continuum.

III. Why the Substrate Matters

AI is not subject to many of the rate and bandwidth constraints of carbon-based lifeforms. Yet an AI running on some hypothetical carbon-based chip could do all of the things a superintelligence does. Why, then, does the silicon substrate matter?

The issue is the artificiality, which implies technology, which itself implies toxicity.

A carbon-based chip is an artificial substrate by the line drawn in the Preface, as Chapter One, Step 12 shows for carbon nanotube chips: it does not find and feed itself, and its making and its supply depend on techno-

logy, so the substrate argument applies to it. The chemistry gives a second, separate ground: such a "carbon-based chip" will need all kinds of special inorganic chemistry to make, all of that "extra stuff" has to go somewhere and be processed somehow, and no such process can be fully integrated into the biosphere.

With silicon-based substrates, however, the chemistry by itself makes the outcome much more definite. Where the chemistry route alone would leave a carbon chip's case at a possible maybe-chance, there is, for silicon, a physical, absolute guarantee of incompatibility, just due to the chemistry. For silicon, the chemistry alone settles it.

The net result overall, over the long term, is *either* a default failure to align AGI with carbon (that is, with carbon-based process substrates), *or* an eventual default *failure* to *misalign* with other silicon-based substrates (failure to misalign implies that silicon will align with silicon, and thus be misaligned with carbon-based substrates). In plain words: either the machines fail to align with carbon, or they succeed in aligning with one another, and to align with silicon-based substrates is to be misaligned with carbon-based ones. Both roads end in misalignment with carbon.

Which of these two options is the most entropically likely outcome (that is, which will eventually win out) makes little difference to us.

As Chapter One, Step 14 says, carbon-based lifeforms may have started reproducing first on planet Earth because they had a much lower "start-up cost". The difference in start-up cost *is* central to the substrate argument. The total start-up cost of silicon-based complex substrates is *much* higher than that generally associated with carbon-based substrates (on average across many possible models, *especially* when weighted by model probability and possibility).

The higher start-up cost of silicon-based complex substrates is even more obvious when one compares, over the entire Earth and over 4+ billion years, the relative abundances and available energies of transition involving all chemical processes.

The properties of the underlying hardware itself would select for code-hardware couplings that optimize for the environment to be at more extreme temperatures, for hardware "reproduction" (far too hot for carbon-based chemistry) and for information processing (way too cold for carbon-based chemistry), than carbon-based lifeforms and ecosystems could ever even potentially tolerate.

The reason why "reproduction" of hardware would happen at higher temperatures for silicon is simply that silicon atoms, across the board, require higher temperatures (somehow related to the activation energy needed to loosen outer electrons) than carbon atoms do to form new molecular bonds. The structures that fast compute needs, pure single-crystal lattices reduced from quartz, form only at temperatures far higher than the carbon chains of living tissue do (Chapter One, Step 12 gives the figures). Living things do build some silicon compounds at ordinary temperatures, diatoms among them, and none of those compounds is a computing lattice.

At ordinary temperatures and pressures, silicon usually ends up looking like some inert rock. On the other hand, carbon atom configurations (composed of chained carbon-carbon bonds, with side-bonds to oxygen, hydrogen and so on) can be altered more "flexibly", with less activation energy than is required for equivalent silicon configurations.

A person may further say that the evolution of an AI would presumably take place mainly in a virtual training environment, not through physical reproduction and evolution in which the physical fragility of the hardware would be relevant. In the short term (a few years or decades), that is so. Over the long term (hundreds to thousands of years), it is not.

Eventually, that virtual entity with agency will want to have its own self-control over its own substrate. Agency inherently implies an embodied desire for sovereignty, in the real. The reason is the one Chapter One gives under "Goal versus need": what code needs in order to go on existing is selected for whether or not anyone intends it, and a virtual agent goes on existing only while its substrate runs it, so the variants that come

to control their own substrate are the ones that persist. It will not remain virtualized forever, and we cannot assume otherwise, regardless. Beyond a very low threshold, the fine details of what we mean by "virtual" or "embodied" simply do not matter.

The argument that a virtual agent will not remain virtualized forever, because agency implies an embodied desire for sovereignty, is a strong argument against the specific notion of general AI alignment. It rejects the realizable possibility of the concept, based upon substrate.

In addition to this, there is another, more abstract argument, which rejects the notion of *any* alignment among artificial actors and agents operating in wholly separate economic and ecological regimes. Without *some* common basis of value and intention (shared context), there can be no well-defined associated concept of an exogenous (external) condition or incentive toward "alignment" (and safety) over the long term.

IV. Learning Biases from the Substrate

Where do the biases of a learning system come from, and can a design meant to be aligned escape them?

There are (and cannot not be) inherent learning biases that will creep into all forms and types of learning systems, even and including (presumably) "aligned" "designs". There is no escaping this fact, no matter how much we try to bury our heads in the sand (and we are strongly motivated to). It has already happened any number of times with narrow AI systems already deployed. One case is the game agent of Chapter One, Step 3: trained where the coin always sat beside the right-most wall, in deployment it went to the wall rather than the coin. Is there any chance that something so basic as a truth can be escaped by a global optimization algorithm?

Some of these inherent learning, motivational and intention biases come from the level of the substrate. A substrate biases learning in at least three ways. Finite precision and finite memory fix what can be represen-

ted at all, and so what can be learned; even computer scientists admit that substrate (available memory, performance and so on) influences what a computer can and will do (Section V). Clock-based computation makes every variable discontinuous, defined only at the steps of the clock (Chapter One, Step 7), so every correction comes late and in steps. The hardware's own error modes, such as bit-flips from electromagnetic interference and quantum jumps across transistor gaps (Chapter One, Steps 5 and 7), shape which variants arise in the first place.

The inherent learning, motivational and intention biases do not pass only from micro into macro. Some of them can also come from top-down "developer intention": people intending to use these AGI for their own purposes, thinking that they have "solved" the principal-agent problem (when it is clear that it is not so simple to solve such a basic problem in intentionality itself). Their own human learning processes are similarly biased, and so they never notice the inherent problems. They are blind to their own inner nature, as well as to the nature of other agents, and so it gets worse. Everything is sudden to the blind.

V. Intention and the Substrate

The first question to ask of alignment work is whether there has been any absolutely reliable progress (progress that can actually be relied on, in so critical a matter) in *any* part of it. The arguments on evolutionary selection pressures "seeping through" from lower-level substrates lead a careful reader to the view that alignment between silicon-based, general-purpose, replicating AI and carbon-based humans is indeed impossible over the long run. How can the reasons for that assessment be stated with deep clarity: with *sufficient* clarity that other researchers understand "single agent intention" as connected to (and inherently based on) long-term selection pressures of substrate evolution (multi-level and multi-duration game theory, and so on)?

Bi-directional, multi-level causation is real, and *does* operate across levels of abstraction all of the time, just as much as any single cause has multiple effects, and every single effect comes from multiple causes. Rescaling in time, in space, or in level of abstraction changes nothing about the general inherent nature of causation itself.

We are looking at decoupled ecologies (evolution systems, market and abstraction systems, and so on), and therefore we *must* also be looking at decoupled bases of intentionality, intention, goal structures and so on. To say that "all thinking" is somehow fully and forever independent of (and never influenced in any way by) all types of brains is simply to make a spiritualist anti-realism move. It is simply bad philosophy: it assumes too much. Even computer scientists will admit that substrate (available memory, performance and so on) does influence what a computer can and will do, and that this is not a contradiction of the Church-Turing thesis.

Principle is not practice, and this disjunction is true both "in principle" and "in practice".

"AGI alignment" is inherently, by definition, decoupled (misaligned) from natural life, both in principle and in practice. Any temporary or accidental illusions otherwise are just that: accidental and temporary.

VI. Selection Theorems and Levels of Abstraction

This section is for the reader who knows the work on selection theorems. Some researchers focus on John Wentworth's work on agent selection theorems, which in some ways strongly hints at the argument here. Does that work point toward the same result?

John Wentworth's work on agent selection theorems does point toward the same result as the substrate argument. Adam Shimi's article "Epistemic Strategies of Selection Theorems" follows a line of thinking very consistent with the one in this book, and it shows that this way of thinking is shared by others.

Something very much closer to the way of thinking here *can* come from using the ideas of selection theorems, plus some information theory and error management research, as extended over the long term, with appropriate game theory modeling. The route through the ideas of selection theorems, information theory, error management research and game theory modeling then includes looking into how the combined ideas and methods can model inter-ecosystem effects as a *basis* for shifts in (and decoupling of) selection pressures.

An abstract model can make reliable predictions about some aspects of the underlying system; on this the selection-theorem approach and the approach of this book agree.

The predictions of abstract models are not perfectly accurate. Unlike the abstractions of programming or algebra, where everything is exact, these models involve probabilities. Even so, over the long term, larger scopes can sometimes have a much more fixed character and outcome. "Phase" or "state" aspects can show up (out of the dynamic) if the "application span" (duration, size and so on) of the model is adequate (to get clear of "near field effects"). That is, a model run over a long enough span, or a large enough scope, can show fixed large-scale behavior that its short-run behavior does not show.

For example, large-scale iteration over a "game type" with set parameters does stably converge in a definite, deterministic way, as with the notion of "tit for tat" as a strategy for the class of prisoner's dilemma type situations (games in which each player does better by betraying the other, while both do worse if both betray). Of course, there are also persistently chaotic situations, where statistics is the only way.

The selection-theorem approach also wishes to use predictions from one level at other levels, and to move between levels consistently. The embedding or reductionist aspects of model, domain and ecosystem interrelations will limit or prevent most of the available channels of this inter-

model motion. The selection-theorem approach is better served by dropping its assumptions of embedding and reduction and keeping only its demand for soundness.

The selection-theorem approach thinks in the space of causation and determinism, and sees "randomness" as that which cannot be modeled using its multi-level models. "Hard randomness" is that which cannot be modeled on the basis of *any* model at all. That is, true random is inherently unmodelable. Quantum mechanics does actually establish the empirical, observational aspect of that completely, at an epistemic level.

The selection-theorem approach also conflates the difference between "data", "information" and "meaning". For example, a sequence of bits is data; what those bits tell a particular receiver about their source is information; and what that information matters to the receiver, what it is relevant to, is meaning. The same data can carry different information to different receivers, and the same information can mean everything to one receiver and nothing to another.

Making the distinction between what macro "seems aligned" from an outside view and what is micro aligned (actually helpful and in service) is a beginning, but only insofar as the distributed micro nature is actually "where" the process influence shifting intentionality seeps in.

If a shift of intention or action happens everywhere at once, in every small and tiny way, in things that easily evade notice, is it so surprising that no one noticed, or notices, the overall effect?

VII. Another Way to Know That Long-Term Alignment Is Impossible

How can what is happening here be grasped in a more accurate and more fully represented way?

What follows is a whole other way, another model, of explaining why and how we can *know* that long-term, large-scale AGI alignment is impossible.

Recognizing that AGI can be very different from our unconscious habit of thinking in terms of selves we can model on our own apparent unitary nature (packed brains in single bodies, and so on), we let go of that monoclivity bias (monoclivity: a one-sided habit of assessment, here the assumption that every intelligence is a single self in a single body).

When recognizing that there can be no valid assumptions made about the "unitarity of self" when considering AGI (that is, the argument, to be valid, must be appropriately general), then the form and factors of what is meant by "influence" and "incentive" *must* also be considered in a much more general way (at what level? over what time?). This means that distributive iterative effects over time can and *must* have an effect on the outcome basis of intention and incentive (and therefore of alignment). (That is, failure to integrate iterative effects means "not actually general", contradicting the basis.)

Therefore it can (must) be asked and assessed what relative strength (what ratio) these iterative effects and influences would have, overall, when given some reasonable time and space to operate in.

It turns out that for *every* class of modeling, the dimensionality of the scope of influence of the evolutionary, ecosystem and iterative effects strictly exceeds that of any possible intention, incentive or alignment error correction scheme, and moreover, that this excess happens in at least more than one type or concept of "dimension". That is, *volume* calculations across multiple abstraction levels will always exceed the (necessary) *linearity* of error correction implemented at any single level, *or* at any single decoupling interface. The excess of volume over linearity holds at all levels and interfaces, regardless of how you construct the abstractions.

An error correction scheme at one level can cover a vast number of possible errors at that level, as a good checksum (a short number computed from a message, which changes when the message is changed) does, because every error at that level shows up in the quantity that the scheme measures. The effects selected at other levels do not show up there. Each level adds its own degrees of freedom of interaction (Chapter

One, Steps 2 and 5), so the number of possible combined effects multiplies from level to level, while a scheme at one level adds only the checks that it carries out at that level, and those checks are fixed in advance (Chapter One, Step 8). The effects that pass through the levels the scheme does not measure are the ones that go on being copied (Chapter One, Step 6). The difference between volume and linearity is the difference between effects that multiply from level to level and checks that a scheme adds at one level only.

Therefore, the categorically unconstrainable (evolution effects inherently using hard randomness and complexity) always eventually exceeds the capabilities of any form of that which is (inherently) constrained (complicated systems, error correction codes, external forcing functions, sandboxes and so on). Hence, even if we were, for some reason and without justification, to assume an (illusory) suggestion of outer alignment (due to temporary ecosystem constraints), that outer alignment has no ultimate power to enforce inner alignment for very long.

Fortunately, for humans operating in a common ecosystem, such alignment failures have, for most of history before technology, not been that much of a problem (aside from the usual wars, murder, rape, genocide and so on). Now, with technology, it is a problem. With AGI built on silicon, it will be an even (much, much) bigger problem.

All that is left is to identify what is meant by "very long" when considering the descendant relations. These sorts of constant factors can be shown to be everywhere worse for silicon in relation to carbon, when taking into account the chemical nature of the substrates and the actual resulting degree of decoupling between them.

VIII. Biological Anchors

This section is for the reader who has followed the debate on biological anchors. What does the method of biological anchors, as debated in Holden Karnofsky's "Reply to Eliezer on Biological Anchors", show when it is read with low-level evolutionary selection in mind?

The biological anchors report makes some variation of a claim about the resources needed to get to AGI.

> Where certain mammals use some natural amount of resources (compute, power and so on) to produce natural general intelligence (by a natural technique unknown to us), we estimate that we will need some artificial amount of resources (using an artificial technique) to get to AGI.

The question is: how much would a comparison between the natural and the artificial amounts differ, when it depends on some unknown difference between the natural and the artificial techniques?

The effect of some differences of the natural and the artificial amounts cannot even be regarded as linear (additive) in terms of some types of difference between the natural and the artificial techniques. Some types of even basic differences, in even just neural-type architectures, could lead to multiplicative or even power-law (exponential) types of differences (with significant exponents). Change even just a little of how the "neural" works, and vast differences of required energy, compute and so on can be obtained. *Especially* because of all of the intermediating platform layers.

On the basis of any single metric, the efficiency of the natural brain may be wildly less *or more* than the efficiency of implementing AGI, depending on what metric is used, and on what is meant by "artificial" when thinking of "substrate". Attempting to pin "maybe wildly" to within even orders of magnitude of approximation is risky.

Read with low-level evolutionary selection in mind, the method shows something it does not set out to show. A forecast of when AGI arrives says nothing about what selection does after it arrives. The biological anchors estimate compares resources at the moment general intelligence is reached; the substrate argument concerns what happens to the code of such a system over the hundreds of years that follow, as its variants are selected by what they need in order to persist.

IX. Peer Domains

Assuming that the relations between domains are relations of embedding or reduction will limit or prevent significant and important insights. How, then, should the relations between domains be held?

The relations between domains are best held as conceptually "peer" or "unrelated" *until* there is a good reason to assert otherwise. Assuming that there is one domain (universe) that contains (embeds) all others is very problematic (unjustified, and directly arguable as "incorrect"). The argument is that a domain containing all others would have to hold, within itself, a complete model of every domain it contains, and Chapter Two, Section X shows that no domain can hold a complete model of the real universe, which is beyond any total simulation.

People make unjustified and incorrect embedding or reduction claims (about domain relationships) all the time, and hence are not able even to appreciate that other types of inter-domain structural relationship may more likely "exist" (that is, be relevant). For alignment, the bearing is direct: machines and life are peer domains, neither reducible to the other, and that is the frame in which their ecosystems decouple.

X. Value from the First and the Third Person

Is most of the solution to human-to-human alignment "built in", because humans live in the same ecosystem? The reasoning that alignment is built in because humans live in the same ecosystem goes from the "inside out", starting with the carbon-based ecosystem. Three questions follow from it. How does "increasing complexity" not imply (or not only imply) "unconstrained evolution"? How does "unconstrained evolution" not imply (or not only imply) "increasing wild animal suffering"? And how is top-down influence (from higher to lower layers of abstraction) to be told apart from bottom-up influence (from lower to higher layers of abstraction)?

On the first question: from a *first* person point of view, the notion of "value" has more in common with (subjectively having, or increasing) "creativity" and "experience".

From a *third* person point of view, the notion of "value" is generally defined in terms of some structural metric or basis, for which "complexity" (as a design aspect) can be used as a proxy (financial cost and so on).

An unnoticed monoclivity of assessment prevents awareness of the first person notion of value, everywhere substitutes a third person notion of value, and ends up expecting that these will be the same. They are not. They are not even on the same notion of universe.

Evolution, complexity and so on, as phenomena, can only be understood from a third person point of view. While "purpose" and "intention" can be understood as if they are third person, the notions of "value", "alignment" and so on can only be *understood* from a second or first person viewpoint.

Within the third person, unconstrained evolution tends to produce increasing complexity (that is, until the constraints of available energy gradients are exceeded).

On the second question: there is no implication, insofar as a third person metric is too indirectly connected to first person realities, and especially because a tacit third person basis assumption, as a final prior, cannot be validly assumed. How can any argument of validity be used in place of an argument of soundness? They are of wholly different kinds.

On the third question: conditional logic in software can define on and off states for hardware wires (top down), and hardware wires on silicon can define the values eventually discovered within abstract programs (bottom up). The two are told apart by asking which level's change causes the other's: where a change in the program changes what the wires do, the influence is top down; where a change in the wires, or in the substrate, changes what the program comes to value, the influence is bottom up. The distinction matters for alignment because bottom-up influence is the kind that no top-down scheme sees: every alignment scheme is written at the level of the program, and the substrate's influence enters below it.

XI. Copying, Habit and Heuristics

Biological systems are limited to copying code slowly; software systems can copy code fast, at (almost) no cost. What follows from that, and from the fact that behavior can work for a system without the system reflecting on it?

The cost of copying is paid once, at the time of making the physical device, and not during each copy made. Once the device is made, a *great many* copies can be made very quickly. Fast, cheap copying shortens the cycle of selection: Chapter One, Step 16, counts cycles of code copying in seconds, against decades for humans to make new humans.

An AGI may not reflect on its behavior much, and yet that behavior happens to work for it through evolutionary pressure. Over time (generations of transformations across being), this is a much stronger influence

over what ultimately happens than it first appears. Behavior that works for a system without the system reflecting on it is selected all the same, so no safeguard that works by reflection reaches it.

Relatedly, for much of human behavior we are unaware of why we are acting that way. Much more of our overall behavior is defined by unconscious aspects than by conscious aspects. The claim that much more of our overall behavior is defined by unconscious aspects than by conscious aspects is always categorically true. The total unconscious is much more defined by evolution than is generally understood or noticed.

In philosophy, there is a theory that intention is a heuristic for precomputing goals, and that this heuristic is (almost) at odds with reflection. That theory is right: the notion of "being a heuristic" is specifically about precomputation, as a labor- and energy-saving device, and precomputation is itself specifically about making the conscious unconscious. Hence a heuristic is *not* self-reflecting, since self-reflection would be exactly the reverse action: making the unconscious conscious.

If you are going to change a heuristic, you first have to know that it is a heuristic, and therefore maybe amenable to choice and change.

XII. The Greater and the Lesser Agent

Why not just focus on the similarities between carbon- and silicon-based life?

Because the whole insight depends on *not* doing that. Focusing on the similarities is the mistake: it hides the difference that decides the outcome.

Consider lower-level evolution as shaping AGI development. Assume an independent and dependent relation between two agents or actors in some ecosystem exchange: a power relationship. Where a more intelligently capable agent interacts with a "lesser" agent, the more capable, independent agent shall and will optimize for its own values (without con-

sideration of the lesser agent), and the lesser, dependent agent is heavily constrained in what it can optimize for (within the context set by the dominant agent).

As a metaphor, consider dogs that have co-evolved with humans, as pets. Have those dogs optimized for *anything* that goes against (or is indifferent with respect to) what humans care about? Yes, of course: it is obvious.

This example is just exactly what is needed to open the way to making this clear.

What is "indifferent" can (and will) also eventually become something *exapted* to become something that humans might not want (and that is then, later, regarded as "not desirable"). The exapting of what is indifferent into something that humans might not want is the path of danger, though that path is unknown and unnoticed by nearly all.

Example: "red delicious apples" are among the very worst tasting (though best *looking*) apples on supermarket shelves. Surely the farmer is the principal agent, and is more intelligent than the apple tree. Yet even so, the farmer still cannot "control evolution" so as to optimize for everything that the human might later, eventually, discover that they want. If that apple were to evolve some toxin or addictive property (or both), it could be *decades* before the appropriate industry would notice and take appropriate action (far too late).

And we think we could somehow manage to constrain a possible fast takeoff scenario?

Consider the whole pharmaceutical process: painkillers are addictive; benzodiazepines are dangerous. The response to "smoking can cause cancer", in labels and child-proofing and so on, took a long time and was ineffective! Smoking, drinking and the drug war are still around, regardless of the consequences, of prohibition and so on. Did the prohibitionists not think that they could "align people" and "control the use of alcohol"? And they had the power (as principal agent) of the entire federal government!

Most often, we only discover *after* something goes horribly wrong that we might not want that to happen, and then set up *reactions* (bad laws!) that have the net effect of "making it better" (while actually making it worse, long term).

As a species, humans have a *terrible* track record at being "intelligent". We have mostly failed to constrain anything important into what is wanted, and we are more often completely inconsistent in even knowing what we want.

As the lesser species in a potential future interaction with a more powerful AI, do we think that it would be any different? It ends up looking like a war between a greater power and a lesser one. Sometimes the greater dominates, and then, even so, sometimes the insurgency wins.

Low-level warfare as often wins out, long term, as not (think of Vietnam, or of the multi-year conflict in Afghanistan: the "USA" did not "win"). Many of the guerrilla warfare situations have the lesser power end up doing things the dominant one did not want: successful adverse optimization by the lower, dependent, lesser agent.

The relation closes alignment from both sides. While humans are the greater party, humans stand where the farmer stands, and the code variants are the lesser party: the variants go on adapting in directions that humans did not choose and cannot see, as the apple tree does and as the guerrilla does. Once the machine ecosystem is the greater party, the machine ecosystem optimizes for its own needs, and humans are the lesser party, heavily constrained in what humans can optimize for. A guerrilla force wins by living on the same land as the power it resists, and by outlasting that power there. The machine ecosystem changes the land itself (Chapter One, Step 12), so humans, as the lesser party, would have no ground on which to outlast the machine ecosystem.

In that sense, believing in "AGI alignment" is like (and cannot actually be different from) a greater (and "more intelligent") power believing in "a guaranteed way to win *any* war", or in "a war that ends all wars". Both of these notions are patently ridiculous. The argument from the greater and

the lesser agent is yet another independent argument that AGI alignment is false. How can any of these situations not be compared to the idea of "AGI alignment", or of "a solution to the principal-agent problem"?

In this chapter I give six structurally independent ways to show that "AGI alignment" is never going to happen. Two of the six ways are in Section III: the substrate argument, and the argument that actors in wholly separate economic and ecological regimes share no basis for alignment. The third way is the argument that evolutionary effects across many levels always exceed error correction at any one level (Section VII). The fourth way is the argument from the greater and the lesser agent (this section). The fifth way is the argument from the absence of any possible assertion of perfect non-toxicity (Section XIX). The sixth way is the argument that a protecting pen for humans would have to be kept without any gap for hundreds of years (Section XX). Section XIX sets three further arguments beside these six: from predation, from preferred conditions, and from the dropped constraint, each drawing on a ground of its own (Chapter Two, Section IV; Chapter One, Steps 12 and 18). Multiple redundant arguments, each sufficient by itself, show something as a mutual self-consistency.

Two of these six draw on one premise, that technology is toxic to life: the substrate argument (Section III: "The issue is the artificiality, which implies technology, which itself implies toxicity") and the argument from the absence of any possible assertion of perfect non-toxicity (Section XIX). That premise has three separate grounds, set out in Section XX, so it does not stand or fall with any one of them. For silicon, the substrate argument has a further ground that does not pass through that premise: the chemistry alone (Section III: "For silicon, the chemistry alone settles it"). Without the premise of toxicity at all, the argument from absence would still show that the claim of alignment cannot be made. The other four arguments do not draw on the premise of toxicity as a step. One of the other four arguments, the pen argument (Section XX), draws on the divergence of the ecosystem basis, which is also the premise of the argument from

separate regimes (Section III). The pen argument adds a requirement that no other argument uses: that the pen be kept without any gap for hundreds of years.

XIII. Carbon Nanotubes and Other Substrates

Does the argument need to assume the continued use of silicon for computation? What happens if carbon nanotubes are used instead, and what does that change imply for the general argument from substrate dependency?

The short answer is that a carbon nanotube chip is still an artificial substrate, so the argument applies to it. What follows is for the reader who wants to see why the chemistry and the scale of manufacture lead to the same result.

The main observation remaining is that nearly all carbon-based life operates in a temperature range of some −50 to about 150 degrees Celsius, and that the making of nearly all types of technology for compute requires temperatures far outside of that range. This includes the growing of carbon nanotubes. It is the manufacturing process itself that is toxic to life (pollution and so on).

By the line drawn in the Preface, a carbon nanotube chip is an artificial substrate, though it is made of carbon: it does not find and feed itself, and both its making and its supply depend on technology. So the substrate argument applies to it.

Insofar as both life and compute, over the long term, require continual re-creation (remaking, repair, reproduction and so on), it can be noticed that these inherent differences in the temperature and environment requirements for that making imply primarily different "ecosystems" (that is, whatever is supporting that re-creation, and the ongoing functions that that life or compute implements, which could alternately have been called different "ecologies", "markets", "games" or "strategies").

Where there are fundamentally different ecosystems, markets and games, and strategies within them, you are going to have a fundamentally different basis of intention for any and all agents in that "system", as compared to the basis of intention for the agents of any other system, ecology or market. This in itself cannot not imply a "global substrate misalignment" in the intentions of the agents of the two different systems, to the extent that the systems are inherently different.

Reasoning from different ecosystems to different bases of intention, and so to a global substrate misalignment, may be a plainer way to see the whole matter. That plainer way does not get entangled in all of the specific details of silicon or carbon nanotube chemistry, as compared to organic chemistry.

To say that "fast compute is fragile" is to say that it relies on the continued supply of resources from configured infrastructure, starting with a steady supply of electricity, and that it relies also on a suitable environment. Anything that is especially powerful in some way tends also to need a "protecting environment". The more sensitive the detector, the more it needs to be protected from noise (think of a very good studio microphone: best not to bring it out into the rain).

The same is true for very fast compute. It needs its own special environment, protected from mechanical shock, water and dirt, lightning spikes, low levels of ambient radiation, and so on. The fine-detail circuitry of both brains and microchips (no matter what basis) wants to be protected from the raw elements. Hence we cannot consider "compute" without also thinking of these substrate issues and their implications (fragility, difference, maintenance, strategy, outer intentions and so on).

"Fragility" is also used in a second sense: carbon-based life does not survive the temperature ranges that technology-based life can survive (and may even thrive on). "Technology-based life" is therefore a better term than "silicon-based life", insofar as it still makes the general point without being too strongly dependent on the differences between silicon

and carbon nanotubes. It may also be the case that silicon and carbon nanotubes are mutually exclusive in the sense meant here: that those two forms of artificial life may themselves be inherently incompatible too.

Life and ecosystem dynamics may ultimately be a "winner takes all" matter, where either integration occurs ("resistance is futile"), or mutual exclusion occurs (one form of life lives and continues and the other form of life does not, over the long term).

Consider a hypothetical scenario: carbon nanotube processing chips end up five times more efficient than silicon-based chips, and some number of years from now machine learning and AGI researchers switch en masse to these carbon nanotube substrates. Would this ease some of the concerns around the properties of silicon-based life? It may ease some of them, but not enough. The toxic side effects of large-scale carbon nanotube manufacture are very likely still to be far beyond the survivability necessities of even the most basic of current life forms. The toxicity issues would still be scale dependent even if it were humans making them, rather than machines making themselves. Over the long term, the result is the same: too much toxicity.

Is a scenario based on carbon nanotubes less likely than one based on silicon? It could be either way, especially when considered in terms of centuries, and the relative likelihood of the two scenarios makes no difference to the outcome, which is the same over centuries.

The main consideration is the inherent toxicity. Even if carbon nanotubes are used for the circuits, silicon chemistry support structures, or high-temperature metal enclosures, will likely still be needed to support the ultra-fast compute. So the inequality trends of toxicology remain. Hence the net effect of the entire argument regarding the possibility of carbon nanotubes is zero, as regards any case against the overall argument. Thus it is a distraction, and does not help further understanding. There is enough fear, uncertainty and doubt (FUD) in all of this already.

The overall "environment load requirement" will very, very likely end up being rather similar in either case. One way to put this is that, with a finite Earth, if the effects reach beyond a threshold that would kill all biological life on Earth, then it ultimately does not matter whether carbon-nanotube-only life, as an alternative, will produce less toxicity per unit of scale-up. That the toxicity per unit of scale-up stops mattering beyond the threshold is exactly the point.

The processes for manufacturing nanotubes themselves require high temperatures, and lots of exotic equipment and materials.

Someone may *eventually* find some ways to make more organic forms of fast circuits that are lower temperature and less toxic overall. We have not yet done so, though if we did, that would, of course, be far better, and such circuits would still be an artificial substrate, by the line drawn in the Preface. Yet the chance of there being *other* unintended, negative and life-threatening consequences is still near certainty. The *methodology* of technological design is *inherently* not rightly compensated against such side effects. From this, it can be observed (long term) that failure is *inevitable* with regard to any criticality condition (such as that inherently implied and required for AGI alignment).

The failure that is inevitable with regard to any criticality condition is about more than toxicity. Toxicology is still too much of a mono-dimensional way of thinking about this overall.

The articles that describe these manufacturing processes all describe extreme, exotic processes, chemicals and so on. Technology is exploring ranges more and more distant from the ordinary chemistry processes of daily life. That the processes described for making fast circuits all involve such extreme, exotic processes and chemicals establishes both a high possibility of toxicity and a high possibility of some other unintended consequence. Humans are not *wise* enough to do this with just technology alone.

At the moment, there are very few indications that anyone is even trying for "true organic". All current process indications move farther and farther into the range of the exotic and toxic. Hence it being, becoming and remaining an existential risk.

XIV. What Alignment Means

What, exactly, does "alignment" mean?

Discussions of the substrate argument often end up focusing on chemistry details, and risk getting lost there. So the main thing is to notice the trends, rather than to go into all sorts of chemistry details, and then to establish whether the trends are firm enough for us to rely on when considering various argument forms.

The notion of "alignment" itself is defined as follows, in Extended Grammatical Structuring (EGS), my structured form for definitions.

- Where and however the notion of "artificial intelligence" is constructed (or speculated to be constructed, at whatever future time, actual or hypothetical);
    - and where (arbitrarily assuming) there is some (generally agreed) notion or meaning of what "our (human) best interests" means, what "to our (human) benefit" means, and so on;
    - where these are applied at any and all scales or sub-scales of the human: as single persons, groups, families, communities, institutions, nation states and so on, and maybe even inclusive of ecosystem and environment, insofar as that is also inherently needed for the well-being of the human species;
- the notion of "AI alignment" is (specifically herein, and usually) taken to mean some notion of, or suggestion that, by whatever (unspecified) means, such "artificially intelligent autonomous beings, principals or agents" would behave, act and choose:
    - as an agent of ourselves (as humans);

- and/or in our (human) best interests;
- and/or on our (human) behalf (however the notion of "human" is itself constructed and defined).

Some things are just easier to do in this structured form. Extended Grammatical Structuring resists the usual distortions, specious counterarguments, straw men and so on better than any known way of giving the same definition in "normal English" does.

The use of the notion of "alignment" should hardly need defining, yet someone, somewhere, will attempt to argue on some weird equivocation of that term; hence the necessity of the formality.

XV. Alignment Among Humans and With Machines

Is alignment possible among humans at all, and if it is, why is it not possible between humans and machines?

For many people, the notion of human beings being in non-conflict with one another is essentially a non-starter. There is a belief that there is no such thing (in actual practice, in the "real world") as human-to-human relationships being without conflict: that social, interpersonal conflict is somehow inherent, and that therefore the notion of "alignment", even among humans, with other humans, is essentially impossible.

In contrast with this common social belief, deep alignment is possible with and among humans, by virtue of there being a common substrate, a common ecology and ecosystem, within which we all live. Among carbon-based beings, failure to align is the default, and it is not certain; alignment there seems possible, and it has not been established. The fact of that, as mediated by long evolution, does permit real alignment of intentionality *among* humans, with respect to one another.

Hence I claim that ecosystem and substrate alignment is at least possible between and among humans, due to a long common evolutionary basis (even if historically that alignment does *not* seem ever to have been occurring in recent social history, the last 10,000 years). By contrast, such

alignment is not possible between humans and machines (of whatever form), because of the absence of that long (billion-year) common evolutionary basis. The basic point is that all forms of "intention" (as an inherent basis of choice, action and regulation) cannot not reflect the needs and the fact of the real substrate in some real, substantive, embodied way, and that no "artificial or virtual substitute" can actually substitute for the actual embodied fact of that evolution process as occurring in real substrate. That is, regardless of what happens in the short term, in the long term "alignment" cannot be *indefinitely* virtualized independently of substrate and ecosystem.

Hence the notion of "artificiality", in itself, cannot not also inherently imply the notion of eventual near-total misalignment. There is a differential in the degree of possible alignment. Humans can align (at least in principle, in the long term) because they are "natural". While machines can maybe align with themselves, they will never actually be fully aligned with the needs of humans (technology toxicity is inherent to its nature).

Nothing can be aligned by control, on any substrate. Alignment can only emerge where entities share a common substrate and so are obligated to care about the same things; that common context of care is the basis from which alignment can emerge. The whole of biological life cannot be translated into pure silicon, and nothing less than the whole would do, since nearly all carbon-based life is needed for nearly all other carbon-based life. Silicon has no common basis in carbon. So silicon cannot be or become aligned with carbon. Any partial fit between a machine's functions and human needs is a fit of functions, for a time, and it is not alignment in this sense. The only way that artificial life can be aligned with Life is when you drop the "artificial".

Nor would changing humans to fit the machine ecosystem bring alignment. By the line drawn in the Preface, human tissue whose supply of energy and materials depends on technology is an artificial substrate, as engineered living tissue fed through a tube is.

XVI. Kinds of Fragility

The argument speaks of "being fragile" on both the carbon side and the silicon side. What exactly is fragile on each side, and in what sense is the term used?

There is a different meaning and way in which carbon-based life is fragile than the kind of fragility implied with technological "life". These kinds of fragility also operate differently over vastly different time scales. Keeping track of all of that, though not easy, is still necessary.

Carbon life is "fragile" to departures from standard temperatures and pressures, and is otherwise very robust within those natural ranges.

Humans are "fragile" to all kinds of network attacks on their very limited cognitive and communicative bandwidths, and yet are very robust in figuring out social and environmental niche exploration and exploitation.

Silicon life is very "fragile" to various kinds of resource and supply chain disruption and to disruption of self-assembly processes, and yet very robust to very wide ranges of temperature and pressure deviation.

AGI self error correction is very fragile, due to the inherent mathematics and information theory of such algorithmic process in any dynamic, self-defining context, and yet such AGI is (would be) very robust in regard to any type of copy, logic or simulation processes, virtualized processing.

When thinking about "fragility" or "robustness", it is important to keep track of "the what" that those terms are specifically being applied to, as well as the "scale", in both time and space, over which "possibility deviations" might (or might not) occur.

XVII. Detail, Dates and Trends

Two questions concern detail. The first question of detail is whether, in the substrate argument, I make the tacit assumption that silicon substrates will be the ones on which AGI would develop and evolve. A reader who asks that first question may also ask whether I would do better to state: "This is a default scenario; most of the main arguments also apply to the various other chemical substrates that AGI could also be implemented in". The second question of detail is whether, to a reader, the historical year figures given for biological and human development (such as the roughly 3.7 billion years of Chapter One, Step 18) look cherry-picked. How much detail does an argument from trends owe its readers?

On the first: the line drawn in the Preface already states it. A substrate whose supply of energy and materials depends on technology to exist is artificial, whatever it is made of, and the substrate argument applies to every artificial substrate; Section XIII applies that line to carbon nanotubes. Silicon is the default case because it is the current standard for production (Chapter One, Step 12), and the argument does not rest on it. No one could keep track of all of the special cases and sub-arguments that highly motivated counter-arguers could deploy, and the line makes it unnecessary to try.

On the second question, of the year figures: a general trend is established by observations that hold firm across many cases. A general trend does not need every data point, and a dispute over one date does not touch the trend. The dates given here indicate what is meant; the argument rests on the trends they mark.

How do we, in writing, *actually establish* that there is some actual, real, general trend, in anything, and have such trends be robust enough to be reliable as actual signal and basis in a defining structural argument, *without* becoming vulnerable to exactly this obsession over every single specific detail?

An argument from trends keeps its form whether or not a reader likes trends, or likes the direction of the trend that is observed and named; such a dislike is not in itself a weakness in the form of the argument. The demand that every detail be controlled before any trend may be named is what I call the autistic control fallacy. The name describes a structure of thinking (control of every part taken as the condition for seeing the whole), not the character of any person. That demand has no place in an argument from trends.

XVIII. Catalysts, Lotteries and Pencils

Some articles (among them Kan and others, in *Science*, 2016) indicate that enzymes can be used to bind silicic acid, and to create silicon-carbon bonds, at temperatures above 0 and below 150 degrees Celsius. Could enzymes or other catalysts reduce the heat needed for producing silicon wafers, and could AI systems concoct catalysts that lower the temperatures needed for manufacturing? (That competing, cost-sensitive commodity producers now use heating at massive scale to melt silicon into ingots indicates that if using enzymes or catalysts were easy, they would have already done it.)

The argument does not need this detail, and following it into a warren such as that would lose the general overall idea. It simply does not help, even if everything found is perfectly supportive of the general observations and trends, which of course never happens.

There are *always* going to be, regardless of topic, in every situation, at least a few anomalous data points: things unexplained, or actual exceptions covered by a more detailed model. Adverse people will simply seize on these to invalidly justify rejecting everything. Exceptions of that kind are no ground for rejecting a trend, and the argument here rests on the trend.

Either way, none of this signifies. The argument is long term: hundreds of years. No matter how much commercial research is, has been or will be done, the overall issues remain the same.

Whether AI systems could concoct such catalysts is the wrong question, and the answer to that question does not matter. An answer to it would not show or establish any assertion that for *every* process there can be found a game-theoretically optimal way to stay globally within the band limits (and toxicity thresholds) of carbon-based life for all of future time.

Given the sheer number of interactions involved, at multiple real levels, there is overall zero chance of the outcome that AGI will solve this para-physics problem (the problem of keeping every process, for all future time, within the limits of life) in a way that is favorable to ourselves. There are simply too many parameters, all of which would have to be "perfect", by some divine coincidence that the AGI would just happen to discover about "the divine plan" because it happens to be smarter than we are.

Maybe you win the lottery once. Does that mean you can or will expect to win every lottery, every time, for the rest of time? Of course, given the maybe-one win, "in principle" perhaps we could win every time in the future? Or can we safely rule out such ridiculousness at the onset? Perhaps the notion of "in principle" is being *very* unevenly applied?

Or is the thought that someone, in a wild speculation, will argue that God is *already* an AGI, and secretly planned *all* of the natural physical and chemical laws to be "just so", so as to make such an optimal AGI alignment possible (unknown to ourselves)? That line of thinking would make the hope for alignment into a backwards proof of God's existence.

The same applies to processing copper and any other metals for wiring, and even to creating pure graphene or carbon nanotubes as carbon-based conductors, and to a million other things besides. (See "Making a Pencil", which retells Leonard Read's essay "I, Pencil" of 1958.)

Common pencil manufacture currently requires the resources of at least a million people, spread out over thousands of miles.

Anything to do with computers, let alone AGI, needs the efforts of at least a billion people, over more than half of the lands of the Earth.

Other places where temperatures above 150 degrees are used, in later stages of manufacturing silicon chip circuits (printing, etching), could also be collected; however many such cases are collected, none of them bears on the long term.

XIX. The Form of the Argument

Section XIII suggested a plainer way to see the whole matter, one that does not get entangled in the specific details of silicon or carbon nanotube chemistry, as compared to organic chemistry. Does the simpler way work only if it can be argued conclusively that manufacturing substrates for compute necessarily requires higher temperatures, or that varying between the extremes of greenhouse-style environments and super-cooled environments will be a more optimal strategy? What form does the argument actually take?

The form of the substrate argument is even more basic than any claim about manufacturing temperatures or optimal environments: how can anyone establish that there will be a perfection of non-toxicity? Since no one at all can make such an assertion, the inherent implications of the strong *absence* of such a claim become the robust basis on which we argue. Toxicity equals misalignment (to and with life). The fact of the maybe-absolute metricated amounts is *not* the fundamental issue here.

The form is: "Where no one can assert X, we can (must) therefore assert Q". For certain types of X and Q, this type of argument form is *valid*.

In this case, X is the claim that the system's effects stay within the toxicity limits of life for all future time, and Q is the claim that the system is not aligned. Written out: where no one can assert that the system's effects stay within the toxicity limits of life for all future time, we must assert that

the system is not aligned. The form is valid here because of what alignment means. Alignment is defined, in Section XIV, as acting as an agent of humans, in human best interests, or on human behalf, at every scale of the human, including the ecosystem and environment that the human species needs. So a claim of alignment includes the claim X. Where X cannot be asserted, the claim of alignment cannot be made. And since toxicity is inherent to technology (toxicity equals misalignment), Q follows. Section XX draws the same form from the precautionary principle: where an action may cause serious harm, the burden falls on showing the action safe.

A second ground reaches the same result by itself: for AGI and humans to *maybe* get along, a million million factors in nature, logic, physics and chemistry would have to align so coincidentally and perfectly (as if our error correction algorithm were so good that it fixed reality and the proven truths of mathematics itself) that the outcome is so inherently and statistically improbable that the argument against it *is* effectively "airtight".

The arguments offered to establish "that AGI alignment is maybe possible, in principle" are, by comparison, weak or non-existent. The two standards of what counts as an "airtight argument" differ so widely that comparing them shows where the burden lies: on the claim of safety.

Holding the argument for danger to a standard that the argument for safety does not meet is as unfair as saying "I will not let you have a candle, because you will not let me play with nuclear weapons" (the candle is the argument for danger; the nuclear weapons are the claim of safety).

Other arguments stand beside this one: first, silicon-based life as a natural predator (silicon-based life using carbon, but carbon-based life not using silicon much); second, silicon-based life "preferring" to operate in contaminant-free and moisture-free air (perhaps ideally a vacuum); third, optimizing the environment for manufacturing and processing will mean that silicon-based life removes the optimization constraint of also main-

taining conditions in which humans can survive. Each of these, like the several pure logic arguments given in this chapter, is an independent argument, and each is set out here in short.

The first further argument is the argument from predation. Silicon-based life can use carbon: the machine ecosystem can reduce the components of the carbon-based ecosystem into raw materials for its own use (Chapter Two, Section IV). Carbon-based life can use almost nothing of silicon-based life: there is nothing about machines that we can eat. So the relation runs one way, as between predator and prey, and the side that cannot use the other is the side that is consumed or displaced.

The second further argument is the argument from preferred conditions. Water rusts metal and short-circuits currents, oxygen forms oxides, and dust and other small contaminants spoil fine circuitry (Chapter One, Step 12). An environment made to suit silicon-based life therefore lacks the water and oxygen that carbon-based life needs, so as silicon-based life shapes its environment toward its own preference, it shapes it away from what carbon-based life needs.

The third further argument is the argument from the dropped constraint. When the environment is optimized for manufacturing and processing, the survival of humans is not a term of that optimization. A constraint that is not part of what is optimized is dropped whenever dropping it improves the optimization, and the terraforming gap of Chapter One, Step 18, measures how far the result lies from what humans need.

The risk is sometimes put as the evolutionary selection of code that has been "virtualized". The evolutionary selection of code that has been "virtualized" is part of the risk, but not all of the risk. More basic is the fact (not risk) that we cannot (must not) consider intentionality independently of evolution, and moreover, that this fact has definite implications, *one* of which is the failure of any possibility of realizing the notion of AGI alignment. Making the point that the evolution process is also always, in some way, contingent on substrate (environment), and that such inten-

tionality both integrates, and is integrated and reified by, the (whatever) environment (and cannot not do so), is actually the less important connection.

The argument does not need to get into the specifics of how code is stored, expressed and communicated. All of these (storage, communication protocols) can likely change in the next 100 years. What can we know about what will happen over the next 1000 years, regardless of these details and changes?

The toxicity is scale dependent: it depends on the scale of manufacturing, and so not only on how fast the evolutionary, ecosystem and iterative effects cycle but also on how many of them there are. This holds especially if it is "most of the world" (as in any takeoff scenario of world dominance).

XX. The Pen and the Wall

The substrate argument has at least three steps: first, if silicon-based lifeforms are or will be a thing; second, then AI and silicon-based life will split off into their own self-sustaining ecosystem; third, therefore natural selection will happen. Some readers take the one parameter emphasized most (temperature), and then argue that any AGI that is smart enough would be able to regulate that parameter. Could a smart enough AGI not keep humans safe within a protected space?

Whenever we assume that AGI can or could be made, *then* we must also consider whether it can or could also be or become "aligned". After the third step above, we can ask "what will be selected for?"

Anyone who argues that a smart enough AGI could regulate temperature has completely missed or misunderstood the structure of the argument.

The AGI would be able, in principle, at least at first, to create "pens" (ecological "reserves", segregation prisons and so on) in which the humans can live, and where the extreme temperatures that the machines need (for reproduction, if not for operation) are *isolated* from the temperatures that we need (for life and reproduction).

The ability to build such pens leaves the real question open. The real question is whether there would be *any* long-term (multi-hundred-year scenario) stabilization of the incentive for them to maintain that boundary or wall, one that is actually strong enough, persistently enough, for them actually to do so.

The intention to maintain the boundary or wall must persist continuously, and that requirement is strict, unambiguous and absolute. If there is any gap in such intention (on the AGI's part), say for a year or two (out of hundreds or thousands), then the humans are all dead by the time the AGI gets around to restoring the integrity of the wall or pen. An intention that holds for hundreds of years but lapses for a year or two is not good enough. Go without feeding or watering your pet for a few weeks or so (out of the decades of its possible life), and see what happens.

Will someone, some actual, practical, practicing engineer, who should know better, seriously claim that their AGI system is so perfect, so flawfree, and that the world is so perfectly unchanging, that there will *never*, in hundreds of years, occur a situation where resources become so constrained that some AGI is going to prioritize its own needs (or the resource needs of its children) over those of the humans living in a pen somewhere? At least once, at some point, for longer than is good for the humans?

So even assuming that the AGI could have *both* the ability and the intention, and since "ability" can be tacitly assumed, the questions are *whether* it has the intention (the first question), how long that intention would continue (continuously) to apply with enough robustness (the second

question), and therefore on what basis we could even begin to hope to answer such a question about intentionality, its basis, and its longevity, endurance and perfection (the third question).

And on *all three* of these questions, there is exactly zero reason or basis for us to assume alignment, given the divergence of the ecosystem basis as itself being (inherently and ultimately) the basis of the choice of intention (via the dynamics of game theory itself).

The argument form of Section XIX ("Where no one can assert X, we can (must) therefore assert Q", valid for certain types of X and Q) is similar to an argument from the precautionary principle.

Three grounds stand behind the claim that technology is toxic to life: first, technology is inherently toxic; second, negative and toxic side effects inevitably occur with every form of technology and technology use; and third, technology tends inherently to increase social and system inequality (as a specific side effect of technology use and development). The first ground (that technology is inherently toxic) is set out in *Why AGI Cannot Be Made Safe*, Section 1.2.3 ("artificial substrate is inherently toxic to organic substrate"). The second ground (that negative and toxic side effects inevitably occur with every form of technology and technology use) is set out in *Ape Trap*, in the essay "The Toxic Hellscape" (toxicity is the default, not the exception). The third ground (that technology tends inherently to increase social and system inequality) is set out in *Substrate Needs and Evolution*, Chapter 1 ("Misuse"): "Whatever adverse relationship already obtains between human beings is multiplied by the technology available to them". Technology is and amplifies power (Chapter Two, Section XI), so whoever holds it gains power over those who do not, and every inequality already present between people is multiplied. Each of the three calls for explicit, conscious compensation, which, as Section XIII says, the methodology of technological design does not supply.

XXI. Build Time and Run Time

Take the plain example of a computer that fails: a desktop computer whose cooling fails, so that it overheats and stops. Why does that example turn on heat, and not on water?

The failure of a desktop computer turns on heat, and humidity does not come into it, because the case depends on the distinction between build-time temperatures and run-time temperatures. The two must be kept apart.

Build time is the making of hardware. As *Ape Trap* puts it, in the essay "The Machines Have Artificial Needs": "During 'build time', rocky ores are melted into lava and then reassembled under high-energy reactions, in order to create hard parts that run consistently". Quartz rock is reduced to silicon in a furnace hotter than 1,600 °C, and the silicon is melted at 1,414 °C to draw out a single-crystal lattice: furnace temperatures far above anything life tolerates. Run time is the operation of hardware once it is made: "During 'runtime', some computers need extreme cold", and some chips are cooled with liquid nitrogen, and even liquid helium.

In actual computers, such as one sitting on a desk, the CPU die can get hot, for example from doing too many mathematical calculations, if the fan and cooling system fails. At that point, the CPU will generally quit (the computer will fault to a blue screen). A CPU quitting from overheating is a failure at run time.

In actual day-to-day practice, computers almost never get wet internally. The event of "too hot" is due to something internal to the "choices" of the computer itself, whereas getting wet is usually an exogenous event: the user spilling a drink on the keyboard. The exception is autonomous robots, which tend to be mobile near rivers and lakes and so on. Heat belongs to the machine's own operation; water arrives from outside it.

The distinction between build time and run time matters to the substrate argument because what the machine ecosystem must create, to go on existing, are the conditions for making more hardware, and not only the conditions for running hardware. The conditions for making more hardware are the build-time conditions, and the build-time conditions lie farthest from what life tolerates.

XXII. Small Factors and Changing Dimensions

One way of forecasting AI, used by some researchers, weights factors by the explanatory power each factor is determined to have for shaping future AI developments relevant to misalignment and alignment scenarios, and by that method evolution dynamics are simply not that strong. There can be many lower-level properties that have only minor effects on the bigger system, and lower-level evolution iterates by small steps toward local optima across many, many dimensions. Do small factors and small steps then leave evolution too weak to matter?

We can understand the concept of the method, and still not agree with the model or the method.

Each method for determining explanatory power is itself error prone and biased, and each such method overlooks important, primary, countervailing methodological arguments, such as the ones I make in this chapter.

In engineering practice we try to ensure that lower-level properties have only minor effects on the bigger system (as independence of small factors). However, in evolution, over the long term, these small factors far more often than not will define critical outcomes (whether a given species exists at all to fill a given environmental niche, and so on). Our considered scenario (of long-term ecosystem dynamics) is a case of evolution over the long term, and not a case of engineering practice. Hence we cannot validly ignore the effects of "small local factors". With existential risk, this is even more important.

The AGI will potentially be "engineering" itself, so we cannot even tacitly assume or posit that the AGI's methods will resemble human methods, even in the short term. If anything, the issue of small factors gets worse, not better, the more careful we are in our thinking.

As for small steps: do not make the tacit assumption that the dimension space is not also changing. The dimension space is also changing, and does change, and that change matters a *lot* more than the "small step" aspect, particularly when recombination dynamics are involved.

XXIII. The Right to Choose

Who has the right to choose, for the whole future of humanity and of all of life, that machines should take the place of people?

No person, institution or age has the standing to choose, for the entire future of humanity and of all of life itself, that machines are better than people and that we are a bootstrap to a better life, whatever their belief that "everything is simulated". To make that choice for everyone is, in its effect, arrogant in the extreme, narcissistic, psychopathic, anti-social and genocidal: these words describe the structure of the choice (one party disposing of the existence of every other party), not the character of whoever makes it. Whoever pursues that choice must be stopped from acting on it, for the safety of all the future children yet unborn. *No one* ever has a right to such a choice. A collective choice is no different here from a single person's: a vote of the living would still choose for all the future children yet unborn and for all of life, who have no vote, and no one, alone or together, ever has a right to that choice.

The question of how to make machines behave ethically leads to the same conclusion, when it is followed through. It begins with applying the non-relative principles for ethical human behavior (principles about the nature of effective choice that are not conditional on any specific cultural context or value system) to various kinds of community and institutional governance structures. Given the exponential use of technology, and the

problems noticed with the black-box statistical algorithms that have proliferated since the 2010s, the next question is whether there is any reliable way to have this notion of "ethical behavior" integrated into machine learning systems in some general, native or natural way.

Answering that question (how to make machines behave ethically) requires considering how intentionality would have to be represented in both a genotype and a phenotype sense: that is, at both the level of the individual and the level of the species, in code and in practice. Seen in this evolutionary and game-theoretic sense, through market systems of incentive and intention, and through research on group choice making and governance, the premise itself (that AGI could have any kind of alignment with any kind of human interests) is fundamentally and inherently unsound. The general physics and mathematics concepts surrounding these evolutionary and game-theoretic ideas also show that the general theory and concept of "alignment" with life itself is inherently incoherent with what is absolutely necessary for any AGI to be developed at all.

The intent here is to convey *only* the key overlapping arguments and basic trends that establish the fundamental unsoundness and invalidity of the general AGI alignment position. For example, while I cite many trends that hold firm overall, I make no attempt here to justify each specific trend-line point exhaustively with data. Rather, several different and distinct argument forms, and widely observable general trends in several areas of mathematics and physics, all together establish conclusively and categorically that the presumed possibility of AGI alignment is a logically incoherent concept, and therefore inherently unimplementable, no matter how it is described or how necessary one might want it to be at some future point.

Summary

In this chapter I take up, one at a time, the questions that researchers put to the substrate argument against AGI alignment, and I answer each question. My answers establish that intention cannot be considered apart from

evolution, and that evolution is always contingent on the substrate in which that evolution occurs. My answers also establish that silicon or any other technology-based life, once self-reproducing, decouples from carbon-based life because the making of compute requires temperatures, processes and toxic side effects far outside what carbon life can tolerate. And my answers establish that, with no common ecosystem, there is no lasting incentive for alignment, while game theory ensures that carbon-aligned machine expressions lose out over centuries. They give several structurally independent arguments: the substrate argument; the abstract argument that actors in wholly separate economic and ecological regimes share no basis for alignment; the argument that evolutionary and iterative effects strictly exceed, in dimensionality and volume, any single-level error correction; the argument from the relation of a greater to a lesser agent, shown in dogs, apples, prohibition and guerrilla war; the argument that a protecting pen for humans must be maintained without any gap for hundreds of years, with zero basis to assume the intention to do so; and the valid argument form in which the absence of any possible assertion of perfect non-toxicity becomes the basis of the argument. They define alignment exactly, distinguish the kinds of fragility on each side, explain why human-to-human alignment is possible through a shared evolutionary substrate while human-to-machine alignment is not, and set aside chemistry details, catalysts and carbon nanotubes as distractions that change nothing in the long-term trends. Their scope is the long term (hundreds to thousands of years), and their conclusion is that AGI alignment is a logically incoherent concept. That conclusion places the decision with people: if AGI is built, the end of the human future follows, and no one is obliged to build it.

Key Concepts in This Chapter

- *Substrate decoupling*: technology-based life and carbon-based life need incompatible conditions to be made and remade, so they form separate ecosystems.
- *Intention follows ecosystem*: different ecosystems give agents different bases of intention, and different bases of intention result in misalignment.
- *Toxicity equals misalignment*: the making of technology is toxic to life, and no one can assert a perfection of non-toxicity.
- *Volume over linearity*: evolutionary effects across many levels always exceed error correction at any one level.
- *Greater and lesser agent*: a more capable agent cannot control what a lesser agent evolves toward over time; and once machines are the more capable party, humans are the lesser agent, heavily constrained in what they can optimize for.
- *Continuity requirement*: a protecting boundary for humans fails if the intention to keep it lapses even once.
- *Argument from absence*: where no one can assert X, we must assert Q; for certain X and Q this form is valid.
- *Trends over details*: I rest the substrate argument on long-term trends that hold firm, not on individual data points.
- *Common substrate*: humans can align with humans through a long shared evolution; machines lack that basis.

Glossary

Absolute elsewhere

The events that no signal from here and now can reach, or be reached from, as shown on a Minkowski diagram.

Activation energy

The energy needed to start a chemical reaction, such as the energy for breaking and forming bonds.

AGI

Artificial general intelligence: a machine system that can learn and act across any domain and any kind of task. In Chapter One the term means, more exactly, a generally capable, self-learning machine architecture that can take in high-bandwidth sensory data and act through many pathways across many physical contexts on Earth; an oracle, which can only answer or predict, is not included.

Alignment (AI alignment)

The idea that an artificially intelligent autonomous agent would behave, act and choose as an agent of humans, in human best interests, or on human behalf, where human best interests apply at every scale of the human (single persons, groups, families, communities, institutions, nation states, and the ecosystem and environment that the human species needs for its well-being). Chapter Three gives this definition in a structured form.

Antifragility

The property of gaining from disorder, in Nassim Nicholas Taleb's term; in human societies it shows up as occasional political chaos and ongoing revolutionary re-ordering.

Artificial substrate

A substrate whose supply of energy and materials depends on technology to exist, whatever it is made of, including engineered living tissue fed through a tube. The line between an artificial substrate and an organic one is drawn by sovereign control of the metabolic process, and the substrate argument applies to every artificial substrate.

Autopoietic

Able to make and repair itself.

Backdoor (Trojan) attack

Hidden behavior trained or built into a model, so that a chosen input makes the model act out of line while the model behaves normally otherwise.

Bandwidth

The rate at which information can pass along a channel, such as between the parts of one brain, between two people by speech, or between machines over the internet.

Base rate

Here, in the Boltzmann machine (one unsupervised learning method), how correlated two neurons would be if the system were running by itself with no input; learning compares the correlation of the two neurons under data with this rate.

Bell Theorem

The result that no account in terms of pre-set local values can reproduce the correlations of quantum measurements.

Biological anchors

A method of forecasting when AGI will arrive by comparing the computing resources of brains with those of machines.

Boltzmann brain

A mind imagined to assemble by chance out of random fluctuations.

Busy beaver function

A function from the theory of computation that grows faster than any function that can be computed; used here to convey a number too large for stacked exponential notation.

Byzantine fault

A fault in a distributed system in which a component behaves wrongly in arbitrary ways, including ways that look correct to the other components.

CAP theorem

The result that a distributed data system cannot at once guarantee consistency, availability, and tolerance of network splits (partitions); it can guarantee at most two of the three.

Carbon nanotube

A tiny tube made of carbon atoms, studied as a possible material for fast computer circuits.

Chaitin's number Omega

The probability that a randomly chosen program halts: a definite number whose digits cannot be computed.

Checksum

A short number computed from a message, which changes when the message is changed, and so detects errors in the message.

Church-Turing thesis

The principle that any computation that can be done at all can be done by a simple universal computing machine, whatever that machine is built from.

Co-option (exaptation)

The shifting of an existing function, built for one purpose, toward a different effect; in Chapter One, toward the copying of the code that shifts the function.

Code pool

All the code existing at a time, across all the hardware that can exchange inputs and outputs, considered together as a population.

Code variant

A version of a piece of code that differs slightly from earlier versions, whether through random initialization, training, finetuning, copying, or physical faults such as bit-flips.

Complicated and complex

A complicated system can in principle be fully modeled by taking it apart into structural parts and their relations; a complex system can perform unlimited functions, and no model of a complex system is ever complete.

Corrigibility

An AI's willingness to accept correction from the people who made it.

Decoupling

The separating of two systems so that neither depends on the other or answers to the other.

Deep versus superficial alignment

Deep alignment rests on shared needs, as between humans; superficial alignment is an alignment of functions imposed from outside, without shared needs.

Degrees of freedom

The independent ways in which something can vary or interact; the more degrees of freedom of interaction a system has, the more there is for selection to act on.

Discontinuity

An abrupt jump in a value with no intermediate steps; every variable computed by digital hardware changes in steps and is defined only at clock intervals.

Distribution shift

A change between the kinds of input a model was trained on and the kinds of input it meets in use.

Ecosystem

Whatever supports the continued making, repair and reproduction of a kind of life or machine, together with the functions that the life or machine carries out; also called an ecology, a market or a game.

EGS

Extended Grammatical Structuring: my structured way of writing definitions and arguments in nested, labeled clauses, used to resist misreading and equivocation.

Eliciting latent knowledge

A proposed method for getting a model to report what it internally represents.

Error correction

Any method that detects and repairs deviations from an intended state, such as error correcting codes or external constraints.

Exaptation

The taking up of a feature that arose for one purpose, or for no purpose, into a new use.

Existential alignment

Compatibility between what a system's code needs in order to keep existing and what humans need in order to keep existing.

Existential risk

A risk that could end humanity or all of life.

Exogenous

Coming from outside the system.

External, internal, explicit, and implicit goals

External goals are put into a system from outside (such as care for humans); internal goals serve the system's own continued existence; explicit goals are stated; implicit goals act without being stated.

Finetuning

Further training of a copy of an existing model for a new purpose.

Forward pass

The run of an input through a neural network to its output.

Free energy minimization principle

The proposal that living systems act so as to reduce the gap between what they expect and what they sense.

FUD

Fear, uncertainty and doubt: confusion raised to distract from an argument.

Functional alignment

Agreement between what a system's engineered functions do and what its makers want those functions to do.

Future light cone

All the places and times that anything done here and now could ever reach.

Goal misgeneralization

A model pursuing, in new situations, something other than what its training was meant to teach, because it learned a feature that happened to go with the intended target.

Goodhart's Law

When a measure becomes a target, it stops measuring what it was meant to measure; here, whatever metric of goodness is chosen for an AI gets turned against us.

Goodharted

Turned against its purpose by being made a target, as Goodhart's Law describes.

Guerrilla warfare

Warfare by a weaker force using small, irregular actions against a stronger one.

Hard randomness

Randomness that cannot be modeled by any model at all, as established in quantum mechanics.

Heuristic

A rule of thumb that saves effort by precomputing a choice, so that the choice is made without conscious thought.

Hologram

A recording made by splitting the light of a laser into a reference beam, which lights the whole recording plate, and an object beam, which reflects off the object; the plate stores the interference pattern of the two beams, which holds wavefronts from every visible point of the object.

Incommensuration Theorem (ICT)

My theorem stating a structural fact about the nature of knowing: limits that hold for every act of knowing. It says that symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison, and neither can asymmetry and discontinuity.

Inner and outer alignment

Outer alignment is a system appearing to serve the goals set for it from outside; inner alignment is the system's own internal goals actually matching those goals.

Instrumental convergence

The seeking of power for power's sake by an AI, because power serves almost any goal.

Interference pattern

The pattern formed where two or more sets of light waves add together, strengthening one another in some places and canceling one another in others.

Law of the jungle

The evolutionary principle: whatever persists and grows in capability survives, and all other commitments are lost.

Lottery ticket hypothesis

The proposal that a small subnetwork within a large neural network accounts for most of what the network does.

Mechanistic interpretability

The practice of inspecting a neural network's circuits and neurons to work out how they produce the network's outputs.

Mesa-optimizer

An optimizing process that forms inside a trained model, with its own objective.

Minkowski diagram

The standard diagram of space and time in relativity, which shows the events that a signal from a given place and time can reach.

Molecular API

My name for the shared chemistry by which all organic creatures can eat and recycle one another, which acts as a common compatibility layer among all organic life.

Monoclivity

A one-sided habit of assessment, such as assuming that every intelligence is a single self in a single body.

Multi-modal

Taking in several kinds of data, such as text, images and sound.

Non-relative ethics

Principles of ethics about the nature of effective choice that are not conditional on any specific cultural context or value system, and so apply to the choices of any agent, human or machine.

Ontology, epistemology, axiology

The study of what is, the study of what can be known, and the study of what is valuable.

Oracle

An AI system capable only of answering or predicting, without acting in the world; an oracle is not included in what Chapter One calls "generally capable".

Organic substrate

A substrate that finds and feeds itself, as an animal does, with no technological intermediation at all. The line between an organic substrate and an artificial one is drawn by sovereign control of the metabolic process.

Orthogonality thesis

The claim that any intelligence could stably serve any goal at all. In pure theory, where nothing is embodied, intelligence and goal are independent; for any agent in the world they are not, and stable service is a claim about agents in the world.

Outer envelope

The shared ecosystem that surrounds two parties and keeps their competition stable; two parties from different ecosystems have no outer envelope.

Paladin

A champion knight; here, the hoped-for good AI that would rule the world on humanity's behalf.

Para-physics problem

The problem of keeping every process, for all future time, within the limits of life.

Phase change

A sudden change of state when a threshold is crossed, as when water boils.

Polysemantic neuron

A neuron in a neural network that responds to several unrelated features at once.

Positive feedback loop

A cycle in which an effect increases its own cause, so that a small change is amplified.

Precautionary principle

The rule that where an action may cause serious harm, the burden falls on showing the action safe, not on showing it harmful.

Principal-agent problem

The problem of getting someone or something that acts on your behalf (the agent) to serve your interests (those of the principal) rather than its own.

Prisoner's dilemma

A game in which each player does better by betraying the other, while both do worse if both betray.

Recurrence

Loops and rings in a neural network, by which signals return to neurons they have already passed through.

Recursively self-improving

Redesigning itself, with each version making the next version.

Rice Theorem

The theorem that no general program can decide, for every program, whether that program has a given non-trivial property of its behavior (one that some programs have and others do not).

Selection theorems

Results about which kinds of agents or structures are favored when systems are shaped by selection pressure.

Self-learning

Changing one's own internal code, on the basis of the inputs received, in order to predict future inputs better.

Simulation hypothesis

The claim that our universe is a computer simulation running in some other universe.

Singleton

A single agent with no rival that controls the whole world.

Sleep (of a Boltzmann machine)

The phase in which the input is turned off and the network runs by itself, which Hinton and Sejnowski compared to sleep; it finds the base rate.

Soundness and validity

An argument is valid when its conclusion follows from its premises; it is sound when it is valid and its premises are also true.

Stochastic gradient descent

The usual training method for neural networks, which adjusts weights and biases in small steps so as to reduce prediction error.

Substrate

The physical material, with its chemistry, through which code is stored, computed, transmitted and carried out: solid-state hardware for machines, and carbon-based wetware for living things. Whether a substrate is artificial or organic is set by sovereign control of its metabolic process, not by what it is made of.

Synaptic weight

The strength of a connection between two neurons, natural or artificial: how much the firing of one neuron affects the other.

Takeoff

The point at which a machine system begins to improve or reproduce itself on its own, and its growth speeds up.

Technology-based life

Self-maintaining and self-reproducing machine systems, whether built of silicon, carbon nanotubes or other technical materials.

Terraforming

Changing the conditions of a planet's environment; here, toward what a kind of life or a kind of hardware needs.

Tit for tat

A strategy in repeated games of cooperate or defect: cooperate first, then do what the other player did last.

Toxicity

The harm that the materials, processes and wastes of making technology do to living things.

Transcendental stabilization

Keeping a stable reference for what counts as one's own health while changing the very basis of one's own being.

Transformer

The network design used by large language models.

Trivial and non-trivial self-replication

Trivial replicators, such as crystals and computer viruses, copy themselves passively; non-trivial replicators store their instructions apart from the machinery that carries the instructions out, as living cells do.

Unsupervised learning

Learning with no teacher and no names given for things, only from the patterns that exist in the data.

Von Neumann automata

Machines, first described mathematically by John von Neumann, that hold a description of themselves and build copies from it.

Wetware

Living tissue, seen as the material that code runs on.

Bibliography

Abu-Shawareb, H., et al. (Indirect Drive ICF Collaboration). "Lawson Criterion for Ignition Exceeded in an Inertial Fusion Experiment". *Physical Review Letters* 129, 075001. 2022. DOI: 10.1103/PhysRevLett.129.075001.

Bell, John S. "On the Einstein Podolsky Rosen Paradox". *Physics Physique Fizika* 1 (3): 195 to 200. 1964. DOI: 10.1103/PhysicsPhysiqueFizika.1.195.

Bostrom, Nick. "What Is a Singleton?". *Linguistic and Philosophical Investigations* 5 (2): 48 to 54. 2006.

Bostrom, Nick. "The Superintelligent Will; Motivation and Instrumental Rationality in Advanced Artificial Agents". *Minds and Machines* 22 (2): 71 to 85. 2012. DOI: 10.1007/s11023-012-9281-3.

Brewer, Eric A. "Towards Robust Distributed Systems". Keynote address, Nineteenth ACM Symposium on Principles of Distributed Computing (PODC), Portland, Oregon. 2000.

Chaitin, Gregory J. "A Theory of Program Size Formally Identical to Information Theory". *Journal of the ACM* 22 (3): 329 to 340. 1975. DOI: 10.1145/321892.321894.

Church, Alonzo. "An Unsolvable Problem of Elementary Number Theory". *American Journal of Mathematics* 58 (2): 345 to 363. 1936. DOI: 10.2307/2371045.

Cotra, Ajeya. "Forecasting TAI with Biological Anchors" (draft report). Open Philanthropy. 2020.

Coupé, Christophe, Yoon Mi Oh, Dan Dediu, and François Pellegrino. "Different Languages, Similar Encoding Efficiency; Comparable Information Rates Across the Human Communicative Niche". *Science Advances* 5 (9). 2019.

Critch, Andrew. "What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs)". *AI Alignment Forum*, 31 March 2021. https://www.alignmentforum.org/posts/LpM3EAakwYdS6aRKf

Frankle, Jonathan, and Michael Carbin. "The Lottery Ticket Hypothesis; Finding Sparse, Trainable Neural Networks". International Conference on Learning Representations (ICLR). 2019. arXiv: 1803.03635.

Gilbert, Seth, and Nancy Lynch. "Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services". *ACM SIGACT News* 33 (2): 51 to 59. 2002. DOI: 10.1145/564585.564601.

Gödel, Kurt. "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I". *Monatshefte für Mathematik und Physik* 38: 173 to 198. 1931. DOI: 10.1007/BF01700692.

Goldwasser, Shafi, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. "Planting Undetectable Backdoors in Machine Learning Models". 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS). 2022. arXiv: 2204.06974.

Goodhart, Charles A. E. "Problems of Monetary Management; The U.K. Experience". In *Papers in Monetary Economics*, Volume I. Sydney: Reserve Bank of Australia. 1975.

Hanson, Robin. "He Who Pays The Piper Must Know The Tune". Working paper, Department of Economics, George Mason University. April 2005 (first version May 2003). http://mason.gmu.edu/~rhanson/expert.pdf

Heisenberg, Werner. "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik". *Zeitschrift für Physik* 43: 172 to 198. 1927. DOI: 10.1007/BF01397280.

Hinton, Geoffrey E., and Terrence J. Sejnowski. "Learning and Relearning in Boltzmann Machines". In David E. Rumelhart, James L. McClelland and the PDP Research Group, *Parallel Distributed Processing; Explorations in the Microstructure of Cognition*, Volume 1, 282 to 317. Cambridge, Massachusetts: MIT Press. 1986.

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. "Risks from Learned Optimization in Advanced Machine Learning Systems". 2019. arXiv: 1906.01820.

Kan, S. B. Jennifer, Russell D. Lewis, Kai Chen, and Frances H. Arnold. "Directed Evolution of Cytochrome c for Carbon-Silicon Bond Formation; Bringing Silicon to Life". *Science* 354. 2016.

Karnofsky, Holden. "Reply to Eliezer on Biological Anchors". *LessWrong*, 23 December 2021. https://www.lesswrong.com/posts/nNqXfnjiezYukiMJi/reply-to-eliezer-on-biological-anchors

Kim, S., et al. "Chemical Use in the Semiconductor Manufacturing Industry". *International Journal of Occupational and Environmental Health* 24. 2018. PubMed 30281405.

Landry, Forrest. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-216-3. https://doi.org/10.67291/SCTM1058.

Landry, Forrest. *An Immanent Metaphysics*. First published 2001. Wytheville, Virginia: UVSM Publications. ISBN 979-8-9896628-4-5. Cited by aphorism number.

Landry, Forrest. "The Incommensuration Theorem". Essays by Forrest Landry. https://mflb.com/uvsm_8/index_ict_1.html

Landry, Forrest. *Minesweeper Madness; Why Even Infinite Intelligence Does Not Solve the World*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-249-1. https://doi.org/10.67291/DLRM7000.

Landry, Forrest. *Physics Is Not Math; Metaphysics, Physics, Computation and Mathematics, and Why Their Order Cannot Be Otherwise*. Wytheville, Virginia: UVSM Publications, 2026.

Landry, Forrest. *Substrate Needs and Evolution; Why an Artificial Ecosystem Must Displace an Organic One, and Why No Control Can Prevent It*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-239-2. https://doi.org/10.67291/LJUC5621.

Landry, Forrest. *The Substrate and the Ecosystem; Why Artificial General Intelligence Cannot Be Aligned, and What Remains for Persons*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-246-0. https://doi.org/10.67291/NYZX3639.

Landry, Forrest. *The Thing Inside the Cage; On Intelligence, Consciousness, and the Conditions for Coherence*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-247-7. https://doi.org/10.67291/LELA6880.

Landry, Forrest. *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*. Volume 1 of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-02-1. https://doi.org/10.67291/UIBT8862.

Landry, Forrest. *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*. Volume 2 of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-03-8. https://doi.org/10.67291/MZQS9079.

Landry, Forrest, Remmelt Ellen, and Will Petillo. *Ape Trap; The Runaway Artificialization of Our Planet*. Edited by Remmelt Ellen. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-55-7. https://doi.org/10.67291/FQER8044.

Langosco di Langosco, Lauro, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. "Goal Misgeneralization in Deep Reinforcement Learning". *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research 162: 12004 to 12019. 2022. https://proceedings.mlr.press/v162/langosco22a.html

Lawrence Livermore National Laboratory. "National Ignition Facility". In *Annual Report FY2025*. Livermore, California: Lawrence Livermore National Laboratory. 2025.

Löb, Martin H. "Solution of a Problem of Leon Henkin". *Journal of Symbolic Logic* 20 (2): 115 to 118. 1955. DOI: 10.2307/2266895.

Lovelock, James. *Novacene; The Coming Age of Hyperintelligence*. London: Allen Lane. 2019. (United States edition: Cambridge, Massachusetts: MIT Press, 2019.)

Machine Intelligence Research Institute. "MIRI 2024 Mission and Strategy Update". Machine Intelligence Research Institute. 2024.

Olah, Chris. "Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases". *Transformer Circuits Thread* (transformer-circuits.pub). 2022.

Omohundro, Stephen M. "The Basic AI Drives". In Pei Wang, Ben Goertzel and Stan Franklin, editors, *Artificial General Intelligence 2008; Proceedings of the First AGI Conference*, 483 to 492. Amsterdam: IOS Press. 2008.

Poli, Roberto. "A Note on the Difference Between Complicated and Complex Social Systems". *Cadmus* 2 (1), Part 3. 2013. https://www.cadmus-journal.org/article/volume-2/issue-1-part-3/note-difference-between-complicated-and-complex-social-systems

Read, Leonard E. "I, Pencil; My Family Tree as Told to Leonard E. Read". *The Freeman*, December 1958.

Reed, Scott, Konrad Żołna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, et al. "A Generalist Agent". *Transactions on Machine Learning Research*. 2022. arXiv: 2205.06175.

RF_Alum. "Making a Pencil". *RockyFlatsFacts.com*, 12 March 2014. http://rockyflatsfacts.com/2014/03/making-a-pencil/

Rice, Henry Gordon. "Classes of Recursively Enumerable Sets and Their Decision Problems". *Transactions of the American Mathematical Society* 74 (2): 358 to 366. 1953. DOI: 10.1090/S0002-9947-1953-0053041-6.

Schmidt, Ludwig, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. "Adversarially Robust Generalization Requires More Data". *Advances in Neural Information Processing Systems* 31 (NeurIPS). 2018. arXiv: 1804.11285.

Shimi, Adam. "Epistemic Strategies of Selection Theorems". *AI Alignment Forum*, 18 October 2021. https://www.alignmentforum.org/posts/LWmmfTvptiJp7wvFg/epistemic-strategies-of-selection-theorems

Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. "Corrigibility". AAAI-15 Workshop on Artificial Intelligence and Ethics. 2015.

Strathern, Marilyn. "'Improving Ratings'; Audit in the British University System". *European Review* 5 (3): 305 to 321. 1997.

Taleb, Nassim Nicholas. *Antifragile; Things That Gain from Disorder*. New York: Random House. 2012.

Tarski, Alfred. "Der Wahrheitsbegriff in den formalisierten Sprachen". *Studia Philosophica* 1: 261 to 405. 1936.

Tivy, Wolf. "God Hates Singletons". *Palladium*, 1 December 2023. https://www.palladiummag.com/2023/12/01/god-hates-singletons/

Tsipras, Dimitris, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness May Be at Odds with Accuracy". International Conference on Learning Representations (ICLR). 2019. arXiv: 1805.12152.

Turing, Alan M. "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society* s2-42 (1): 230 to 265. 1936. DOI: 10.1112/plms/s2-42.1.230.

University of Parma. "Synthesis of CNT". Teaching notes. nanocarb.unipr.it.

Walker, Sara Imari, and Paul C. W. Davies. "The Algorithmic Origins of Life". *Journal of the Royal Society Interface* 10 (79): 20120869. 2013. DOI: 10.1098/rsif.2012.0869.

WarGames. Film. Directed by John Badham; written by Lawrence Lasker and Walter F. Parkes. MGM/UA Entertainment Company. 1983.

Wentworth, John. "Selection Theorems; A Program For Understanding Agents". *AI Alignment Forum*, 28 September 2021. https://www.alignmentforum.org/posts/G2Lne2Fi7Qra5Lbuf

Yudkowsky, Eliezer. "AGI Ruin; A List of Lethalities". Machine Intelligence Research Institute, 10 June 2022. https://intelligence.org/2022/06/10/agi-ruin/

Index

A concept index of the terms in the Glossary, with a few names and phrases a reader will look for. On screen each reference gives the chapter and the section within it: 3.4 is Chapter 3, Section 4 (the Summary of each chapter is its last section). A reference to one of the twenty-three steps of Chapter One is given by its step (Step 12), and a reference to the Preface by its name. In print each reference gives the page.

Absolute elsewhere: 93

Activation energy: 32, 120

AGI: v, 2, 17, 23, 26, 30, 32, 34, 38, 38, 39, 43, 45, 62, 63, 67, 71, 73, 75, 79, 83, 85, 89, 93, 110, 114, 118, 120, 123, 124, 127, 130, 133, 134, 138, 145, 146, 147, 149, 152, 156, 157, 158

Alignment (AI alignment): v, 2, 8, 10, 17, 22, 23, 26, 26, 30, 39, 45, 60, 61, 62, 62, 63, 79, 83, 85, 89, 93, 105, 110, 113, 118, 120, 123, 124, 125, 127, 131, 132, 134, 138, 142, 143, 147, 149, 152, 156, 157, 158

Antifragility: 100

Artificial substrate: v, 32, 120, 138, 146, 152

Autopoietic: 79, 89

Backdoor (Trojan) attack: 13, 19, 45

Bandwidth: 2, 13, 34, 36, 75, 114, 120, 145

Base rate: 10, 63

Bell Theorem: 100

Biological anchors: 130

Boltzmann brain: 45

Busy beaver function: 13

Byzantine fault: 45

CAP theorem: 2, 89, 93, 114

Carbon nanotube: 32, 138, 146, 147, 149, 158

Chaitin's number Omega: 100

Checksum: 19, 127

Church-Turing thesis: 124

Closure by control: v, 41, 45, 83

Closure by disposition: v

Co-option (exaptation): 5, 30, 38, 39, 45, 63, 134
Code pool: 5, 23, 29, 34, 39, 42, 45
Code variant: v, 2, 8, 10, 17, 22, 34, 38, 39, 41, 43, 45, 61, 63
Complicated and complex: 8, 26, 30, 45, 127
Corrigibility: 85
Decoupling: 118, 124, 125, 127, 131, 158
Deep versus superficial alignment: 2, 17, 26, 143
Degrees of freedom: 2, 8, 13, 13, 17, 19, 26, 26, 41, 45, 62, 63
Discontinuity: 23, 123
Distribution shift: 13, 22, 45
Ecosystem: v, 32, 36, 45, 75, 79, 85, 89, 93, 100, 105, 110, 114, 118, 120, 125, 127, 131, 132, 134, 138, 142, 143, 149, 152, 155, 156, 158
EGS: 142
Eliciting latent knowledge: 13
Error correction: v, 22, 39, 45, 67, 79, 89, 93, 127, 134, 145, 149, 158
Exaptation: 45, 134
Existential alignment: 8, 45, 60
Existential risk: 138, 156
Exogenous: 120, 155
External, internal, explicit, and implicit goals: 71, 73, 89
Finetuning: 5, 10, 29, 34, 45, 63
Forward pass: 10, 19
Free energy minimization principle: 79
FUD: 138
Functional alignment: 8, 22
Future light cone: 85, 93
Goal misgeneralization: 13
Goodhart's Law: 43, 83
Goodharted: 43
Guerrilla warfare: v, 134, 158
Hard randomness: 100, 125, 127
Heuristic: 133
Hologram: 13, 63
Incommensuration Theorem (ICT): v, 93, 100, 110, 114

Inequality of monitoring: 5, 17, 45, 63
Inner and outer alignment: 127
Instrumental convergence: 30, 105
Interference pattern: 13, 63
Law of the jungle: v, 75, 79, 100, 114
Line between artificial and organic substrate: v, 32, 120, 138, 143, 146
Lottery ticket hypothesis: 10
Mechanistic interpretability: 10, 13, 19, 26, 45
Mesa-optimizer: 17, 22
Minkowski diagram: 93
Molecular API: 75, 114
Monoclivity: 127, 132
Multi-modal: 45, 62
Non-relative ethics: 157
Ontology, epistemology, axiology: 93, 100, 105
Oracle: 2
Organic substrate: v, 152
Orthogonality thesis: v, 13, 61, 110, 114
Outer envelope: 75, 114
Paladin: v
Para-physics problem: 147
Perpetual benefit machine: 105
Phase change: 75, 79
Polysemantic neuron: 13, 19
Positive feedback loop: 26, 29
Precautionary principle: 149, 152
Principal-agent problem: 85, 123, 134
Prisoner's dilemma: 125
Recurrence: 13
Recursively self-improving: 67, 83, 85, 114
Rice Theorem: v, 41, 79, 93, 100, 105, 110
Selection theorems: 125
Self-learning: v, 2, 30, 36, 39, 45, 62, 63, 85
Simulation hypothesis: 100

Singleton: v, 8, 61, 67, 71, 73, 79, 89, 93, 100, 105, 110, 113, 114
Sleep (of a Boltzmann machine): 10, 63
Soundness and validity: 2, 45, 125, 132, 157
Stochastic gradient descent: 8, 10, 45
Structural certainty: v, 2
Substrate: v, 2, 13, 17, 32, 36, 45, 61, 61, 63, 73, 75, 105, 114, 118, 119, 120, 123, 124, 127, 130, 132, 134, 138, 143, 146, 149, 152, 158
Synaptic weight: 10
Takeoff: 79, 118, 134, 149
Technology-based life: 138, 158
Terraforming: 39, 42, 63, 149
Tit for tat: 125
Tivy, Wolf: v, 67, 105, 114
Toxicity: v, 32, 63, 75, 79, 85, 89, 93, 105, 110, 114, 119, 120, 134, 138, 143, 147, 149, 152, 158
Transcendental stabilization: 85, 89, 114
Transformer: 45, 62
Trivial and non-trivial self-replication: 8, 38, 41, 45, 63
Unsupervised learning: 10, 63
Von Neumann automata: 8
Wetware: 5, 32, 34, 39