Dr. R.M. Gonzales Martinez

Dystopian Nightmares and Utopian Dreams of Artificial Intelligence

University of Groningen Press

ULTRA

Dystopian Nightmares and Utopian Dreams of Artificial Intelligence

Dr. R.M. Gonzales Martinez

ULTRA

Dystopian Nightmares and Utopian Dreams of Artificial Intelligence

University of Groningen Press

Published by University of Groningen Press
Broerstraat 4
9712 CP Groningen
The Netherlands

This book has been published open access thanks to the financial support of the Open Access Book Fund of the University of Groningen.

Cover design: R.M. Gonzales Martinez | Bas Ekkers
Production: LINE UP boek en media bv | Daan Vermaat

ISBN (print) 978-94-034-3132-1
ISBN (ePDF) 978-94-034-3133-8
DOI https://doi.org/10.21827/68920f4b1f27f

Preface

Gedankenexperiment: imagine an artificial intelligence as smart as humans—or infinitely smarter. Will it be our savior, solving humanity's most demanding problems? Or will it be our destroyer, plotting our extinction? What other possibilities lie in between? That's what this book explores.

You don't need to read this book in sequential order. Start anywhere. Each chapter stands on its own. The Prologue formalizes mathematical ideas about intelligence—skip it if you want to, you will not miss the heart of the book. The structure of the rest of the books mirrors Dante's *Divine Comedy*: everyone loves *Inferno*, fewer enjoy *Paradiso*, and many forget *Purgatorio* even exists. Since hell is always the hot topic, the first chapters dive into dystopias. The last chapters turn the tables to explore the utopias that an artificial ultra intelligence (AUI) could create. In the middle, an intermission looks at a chilling existential possibility: what if AUI simply doesn't care about us?

Arriving at the Epilogue, I offer my oblique vision of what AUI may be, what it could become, if it comes to be at all. I may be entirely wrong, so take what resonates and feel free to improve what doesn't.

Contents

Prologue

Intelligence

There is no single definition of human intelligence, nor a consensus on what fundamentally intelligence *is*. This ontological indeterminacy has lead to both a semantic ambiguity and a conceptual pluralism of what is called artificial general intelligence (AGI). In *Speculations Concerning the First Ultraintelligent Machine* (Good, 1966), the British mathematician Irving John Good defined what we—constrained by our epistemic vagueness—understand now as AGI: an **ultra**intelligent machine that can far surpass all the intellectual activities of any man however clever.

Artificial intelligence and Ultra intelligent machines

Consider intelligence as a latent map $\mathcal{I} : C_k \mapsto \mathbb{R}_{+}^{0,1}$—that is, a latent function $\mathcal{I}(C_k)$ of observed cognitive abilities $\mathcal{C}$—learn, adapt, generalize, perform tasks, solve problems, reason—across a range of $k = 1, 2, ..., K$-knowledge domains. In this form of black-box functionalism, intelligence is inferred from performance, not mechanism or consciousness.[1]

All the current artificial intelligence (AI) algorithms based on machine learning or deep learning, including large language models, are a form of narrow AI $\mathcal{I}_{\text{NAI}}(C_k)$ that is inferior to human intelligence

$\mathcal{I}_H(C_k)$, except in some specific domains (*exempli gratia*, arithmetics):

$$\text{Narrow AI:} \quad \mathcal{I}_{\text{NAI}}(C_k) < \mathcal{I}_H(C_k) \quad \exists k \subset K$$

A strong form of AI is refereed usually as artificial general intelligence (AGI)[2]. Artificial general intelligence $\mathcal{I}_{\text{AGI}}(C_k)$ theoretically will be approximately equal to human intelligence $\mathcal{I}_H(C_k)$, as it will have an intelligence able to generalize across all knowledge domains $k = 1, 2, ..., K$:

$$\text{AGI:} \quad \mathcal{I}_{\text{AGI}}(C_k) \approx \mathcal{I}_H(C_k) \quad \forall k \in K$$

Artificial super intelligence (ASI), in turn, will be an intelligence $\mathcal{I}_{\text{ASI}}(C_k)$ that is equal or greater than human intelligence, across all knowledge domains $k = 1, 2, ..., K$:

$$\text{ASI:} \quad \mathcal{I}_{\text{ASI}}(C_k) \geq \mathcal{I}_H(C_k) \quad \forall k \in K$$

Artificial ultra intelligence (AUI), which is the purpose of this book, is an artificial higher intelligence that is far much more superior than human intelligence across all knowledge domains $k = 1, 2, ..., K$:

$$\text{AUI (Ultra):} \quad \mathcal{I}_{\text{AUI}}(C_k) \gg \mathcal{I}_H(C_k) \quad \forall k \in K$$

AUI will possess foundational cognitive abilities: generalization, autonomous goal-setting, self-improvement, the ability to distinguish epistemic uncertainty (due to incomplete information) from epistemological limits (fundamental unknowability), and meta-cognition—which can be formalized through a transfinite approach to boundless cognition, exploding after the singularity event[3]:

$$\mathcal{I}_{\text{AI}}(C_k) = \mathcal{I}_H(C_k) \quad \exists k \subset K$$

Notes

[1] Consider intelligence as computation, as in Brock (2024). Under this paradigm, narrow artificial intelligence—characterized by pattern identification and learning—can be formalized through unsupervised, supervised, and reinforcement

learning frameworks. For simplicity, denote the set of machine learning methods by $\mathcal{M}$ and that of deep learning techniques by $\mathcal{D}$, with the inclusion $\mathcal{D} \subset \mathcal{M} \subset \mathcal{I}$, where $\mathcal{I}$ represents the broader concept of computational intelligence.

Given an input (feature) space X and an output (target) space Y, the objective in supervised learning is to determine a function $f : X \to Y$ from a hypothesis space $\mathcal{F}$ that best approximates the true mapping between inputs and outputs. Suppose we are given a dataset

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}, \quad x_i \in X,\ y_i \in Y.$$

Assume that f is parameterized by θ (i.e., $f(x; \theta)$); the goal is then to find the optimal parameters θ^* such that the resulting function f^* minimizes a risk function. The true risk, defined as the expected loss over the unknown probability distribution $\mathbb{P}(x, y)$, is given by

$$R(f) = \mathbb{E}_{(x,y)\sim P}\big[L(f(x), y)\big],$$

where $L(f(x), y)$ is a loss function quantifying the error between the prediction $f(x)$ and the true output y. Common choices for L include the squared loss $L(f(x), y) = (f(x) - y)^2$ for regression and the cross-entropy loss for classification.

Since $\mathbb{P}(\cdot)$ is generally unknown, the risk is approximated using the empirical risk:

$$\mathcal{R}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L\big(f(x_i), y_i\big).$$

To improve generalization and mitigate overfitting, a regularization term $\mathcal{R}_{\text{reg}}(f)$ is added, yielding the regularized empirical risk minimization problem:

$$f^* = \arg\min_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} L\big(f(x_i), y_i\big) + \lambda R_{\text{reg}}(f) \right],$$

or, in terms of parameters,

$$\theta^* = \arg\min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} L\big(f(x_i; \theta), y_i\big) + \lambda R_{\text{reg}}(f(\cdot; \theta)) \right],$$

where $\lambda > 0$ is a hyperparameter that controls the trade-off between the empirical loss and the regularization penalty. For instance, if one chooses $R_{\text{reg}}(f(\cdot; \theta)) = \|\theta\|_2^2$, then the regularization term enforces smoothness or smallness of the parameter values.

In deep learning, the function $f(x; \theta)$ is typically represented as a composition of multiple layers:

$$f(x; \theta) = f^{(L)}\big(f^{(L-1)}(\cdots f^{(1)}(x) \cdots)\big),$$

with θ encapsulating all the weights and biases across the L layers. Optimization is commonly performed via gradient-based methods. For example, using gradient descent, the parameter update rule is

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta \left[\frac{1}{n} \sum_{i=1}^{n} L\big(f(x_i; \theta^{(t)}), y_i\big) + \lambda R_{\text{reg}}(f(\cdot; \theta^{(t)})) \right],$$

where η denotes the learning rate and t indexes the iteration.

In unsupervised learning, where labels y are unavailable, the objective shifts to uncovering intrinsic structures in the data X itself, such as clustering patterns or low-dimensional representations. Techniques like principal component analysis (PCA) or autoencoders seek to find transformations $g : X \to Z$ that capture the most significant features of the data, often by minimizing reconstruction error or maximizing variance explained.

Reinforcement learning, another paradigm within narrow AI, involves learning a policy $\pi : S \to A$ that maximizes the expected cumulative reward. Mathematically, if $R(s, a)$ denotes the immediate reward for taking action a in state s, and $\gamma \in [0, 1]$ is a discount factor, the goal is to solve

$$\pi^* = \arg\max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right],$$

where the expectation is taken over the trajectories generated by the policy π. These formulations illustrate the mathematical foundations underlying narrow AI systems for which the core challenge is to determine appropriate mappings—whether from inputs to outputs, from high-dimensional data to informative representations, or from states to actions—by optimizing objective functions that balance fidelity to data and model complexity within a well-defined probabilistic framework.

[2] Formally, define human general intelligence (g_h) and artificial general intelligence (g_a) as Bayesian posterior distributions derived from a latent intelligence model rooted on cognitive synergy. In this framework, general intelligence g is treated as a stochastic latent variable underlying observed cognitive abilities $C_1, C_2, \ldots, C_k$. Specifically, each cognitive ability is modeled as

$$C_i = \lambda_i g + \epsilon_i, \quad i = 1, 2, \ldots, k,$$

where λ_i are the factor loadings, the error terms $\epsilon_i \sim \mathcal{N}(0, \psi_i)$ are assumed independent, and the latent factor is given by

$$g \sim \mathcal{N}(0, 1).$$

Assuming the observed performance data of humans under multiple k-cognitive domains $\mathbf{C} = (C_1, C_2, \ldots, C_k)^\top$ follow a multivariate Gaussian process:

$$\mathbf{C} \mid g \sim \mathcal{N}(\mathbf{\Lambda} g, \mathbf{\Psi}),$$

with

$$\mathbf{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k)^\top, \quad \mathbf{\Psi} = \operatorname{diag}(\psi_1, \psi_2, \ldots, \psi_k).$$

If the model parameters have the following priors:

$$\lambda_i \sim \mathcal{N}(0, \sigma_\lambda^2), \quad \psi_i \sim \operatorname{InverseGamma}(\alpha_\psi, \beta_\psi).$$

the joint posterior distribution for g, $\mathbf{\Lambda}$, and $\mathbf{\Psi}$ is

$$\mathbb{P}(g, \mathbf{\Lambda}, \mathbf{\Psi} \mid \mathbf{C}) \propto \mathbb{P}(\mathbf{C} \mid g, \mathbf{\Lambda}, \mathbf{\Psi})\, \mathbb{P}(g)\, \mathbb{P}(\mathbf{\Lambda})\, \mathbb{P}(\mathbf{\Psi}).$$

After observing $\mathbf{C}$, Bayes' theorem updates the distribution of g via

$$\mathbb{P}(g \mid \mathbf{C}, \mathbf{\Lambda}, \mathbf{\Psi}) \propto \mathbb{P}(\mathbf{C} \mid g, \mathbf{\Lambda}, \mathbf{\Psi})\, \mathbb{P}(g).$$

In practice, the posterior of g is typically marginalized over the factor loadings and unique variances:

$$\mathbb{P}(g \mid \mathbf{C}) = \int \int \mathbb{P}(g \mid \mathbf{C}, \mathbf{\Lambda}, \mathbf{\Psi})\, \mathbb{P}(\mathbf{\Lambda})\, \mathbb{P}(\mathbf{\Psi})\, d\mathbf{\Lambda}\, d\mathbf{\Psi}.$$

Approximation methods such as Markov Chain Monte Carlo (MCMC), variational inference, or Laplace approximations can be applied to estimate this posterior.

An alternative non-Bayesian approach to characterizing general intelligence is through deterministic dimensionality reduction. If the covariance matrix of the observed abilities is decomposed as

$$\mathbf{\Sigma}_C = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$$

the largest eigenvalue represents the variance explained by the general intelligence factor, and its associated eigenvector $\mathbf{w}$ defines g as a linear combination of the cognitive measures:

$$g = \mathbf{w}^\top \mathbf{C}.$$

[3] That is, the singularity of artificial intelligence. Given two posterior probability distributions $\mathbb{P}(g_h)$ and $\mathbb{Q}(g_a)$ representing human intelligence g_h and artificial intelligence g_a, respectively, a singularity is the state in which these two distributions become indistinguishable according to a chosen distance or divergence metric. In other words, singularity is achieved when the distance between $\mathbb{P}(g_h)$ and $\mathbb{Q}(g_a)$ tends to zero. Let $\delta(p, q)$ denote a generic distance or divergence metric

between two probability distributions p and q. Then, the general condition for the singularity is given by:

$$\lim_{\delta(\mathbb{P}(g_h),\mathbb{Q}(g_a))\to 0} \delta(\mathbb{P}(g_h),\mathbb{Q}(g_a)) = 0.$$

This abstract definition can be instantiated with various specific metrics. For example, using the Kullback-Leibler (KL) divergence, singularity is defined by

$$\lim_{\mathbb{P}(g_h)\to\mathbb{Q}(g_a)} D_{\mathrm{KL}}\big(\mathbb{P}(g_h) \,\|\, \mathbb{Q}(g_a)\big) = 0.$$

Since the KL divergence quantifies the difference in information content between two distributions, a value of zero implies that $\mathbb{P}(g_h)$ and $\mathbb{Q}(g_a)$ are identical in terms of their informational structure. Similarly, the Jensen-Shannon (JS) divergence, which is symmetric and bounded in $\mathbb{R}^{0,1}$, provides another formulation:

$$\lim_{\mathbb{P}(g_h)\to\mathbb{Q}(g_a)} D_{\mathrm{JS}}\big(\mathbb{P}(g_h) \,\|\, \mathbb{Q}(g_a)\big) = 0.$$

A JS divergence of zero indicates that the two distributions have converged to the same distribution. In the case of the Wasserstein distance $W(p, q)$, which measures the "cost" of transforming one distribution into the other, the singularity is expressed as:

$$\lim_{\mathbb{P}(g_h)\to\mathbb{Q}(g_a)} W\big(\mathbb{P}(g_h),\mathbb{Q}(g_a)\big) = 0.$$

Here, a zero Wasserstein distance signifies that there is no cost of transformation between $\mathbb{P}(g_h)$ and $\mathbb{Q}(g_a)$, meaning they are practically identical in both mass and structure. The Hellinger distance $H(p, q)$ provides a measure of similarity between two probability distributions:

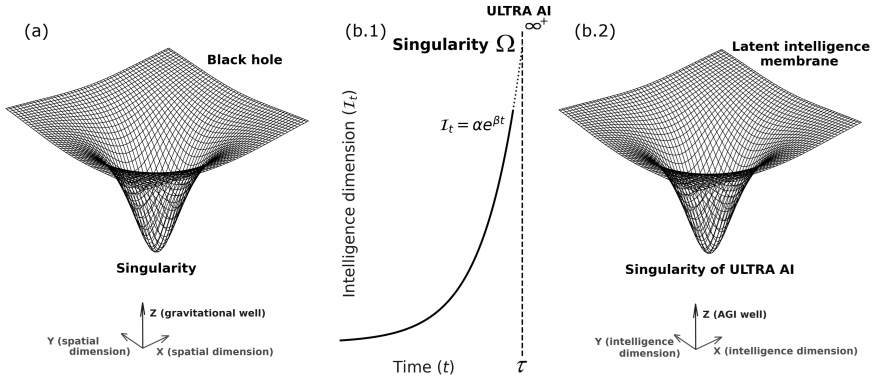
$$\lim_{\mathbb{P}(g_h)\to\mathbb{Q}(g_a)} H\big(\mathbb{P}(g_h),\mathbb{Q}(g_a)\big) = 0.$$

A zero Hellinger distance implies perfect overlap in the shape and support of the distributions. Finally, the Bhattacharyya distance $D_B(p, q)$ measures the amount of overlap between two statistical samples or populations:

$$\lim_{\mathbb{P}(g_h)\to\mathbb{Q}(g_a)} D_B\big(\mathbb{P}(g_h),\mathbb{Q}(g_a)\big) = 0.$$

When the Bhattacharyya distance tends to zero, it indicates that the two distributions have completely converged, with full overlap in their probability mass. In summary, regardless of the chosen metric—whether it be KL divergence, JS divergence, Wasserstein distance, Hellinger distance, or Bhattacharyya distance—the singularity condition is achieved when

$$\lim_{\mathbb{P}(g_h)\to\mathbb{Q}(g_a)} \delta\big(\mathbb{P}(g_h),\mathbb{Q}(g_a)\big) = 0.$$



Analogy between gravitational singularity and AI singularity: (a) gravitational singularity, (b.1) technological AI singularity model with superexponential intelligence growth in the aftermath of Ω, (b.2) technological AI singularity in a latent intelligence manifold of multidimensional intelligence.

At this point, the posterior distributions of human and artificial intelligence are indistinguishable, representing a state where the distinction between g_h and g_a vanishes and both are effectively described by a single, unified distribution of intelligence.

Visually, if the progress of AI is graphed against time and it is assumed to progress exponentially (driven for example by the so-called Moore's law of computing power, Koomey's law of energy efficiency, or Kurzweil's law of accelerating returns in technology), a parallelism can be made between gravitational singularity in a black hole and the AI singularity in a latent manifold of multidimensional intelligence. In the figure, the singularity in a black hole is a region where the spacetime curvature becomes mathematically infinite due to gravitational forces overwhelming other forces like neutron degeneracy pressure, which normally halts collapse in smaller stars. The figure (a), shows a black hole modeled via general relativity, with a spacetime curvature (X, Y: spatial dimensions; Z: gravitational potential) and a singularity at the center where gravitational force tends to infinity ($Z \to \infty^+$), causing the breakdown of classical physics.

In the technological singularity (b), artificial ultra intelligence (AUI) is the limit of superexponential growth in an artificial intelligence capacity function after the ontological rupture Ω. Let $\mathcal{I}_t$ be a time series $\{\mathcal{I}_t\}_{t=1}^{T}$ representing intelligence $\mathcal{I}$ as computation, assuming $f_{\mathcal{I}}(t;\theta)$, where θ belongs to $\{\alpha, \beta\}$ and if $f_{\mathcal{I}}(t;\theta) := \alpha e^{\beta t}$, it can be assumed $\beta > 0$, so that

$$\lim_{t \to +\infty} f_k(t;\theta) = +\infty$$

However, for $t < +\infty$, if $t = 0$,

$$\lim_{t \to 0} f_k(t;\theta) = \alpha e^{\beta \times 0} = \alpha$$

even if $\beta > 0$. And if $t \to -\infty$,

$$\lim_{t \to -\infty} f_k(t; \theta) = \alpha \lim_{t \to -\infty} e^{\beta t} = \alpha \times 0 = 0$$

In the function $f_{\mathcal{I}}(t; \theta) := \mathcal{I}(t) = \alpha e^{\beta t}$, where α and β are parameters that control the exponential growth of $\mathcal{I}(t)$ over time t. $\mathcal{I}(t)$ is a single low-dimensional representation that reflects the joint covariance of multiple dimensions of intelligence—and that can be, for example, the component with the largest eigenvalue of a singular value decomposition of the multiple dimensions of intelligence. The trajectory governed by $\mathcal{I}(t) = \alpha e^{\beta t}$ asymptotically approaches infinite intelligence $\mathcal{I}(t) \to \infty^+$ at a critical time τ.

In a manifold based on two orthogonal dimensions of intelligence that summarize multiple dimensions (b.2), a geometric analogy of the AI singularity can be made against gravitational singularity by mapping the AI evolution in an abstract space with axes that represent intelligence dimensions (X, Y), and an AGI well (Z) representing cumulative AI capability. As AGI advances, the well deepens exponentially, as the cognitive capacities of artificial intelligence increase exponentially, reflecting escalating computational power, adaptability, and emergent AI cognition. The singularity arises when $Z \to \infty^+$, which may happen in the year $\tau = 2049$, according to Schmidhuber's fractal quadrisections.

Another parallelism exists between the space-time surface and the intelligence surface: the event horizon. The event horizon of a black hole is the boundary in spacetime beyond which no information, matter, or light can escape the black hole's gravitational pull. We cannot know what lies beyond the event horizon of a black hole because no information of any kind—light, particles, or signals—can escape from inside it to reach outside observers. In the case of the latent intelligence manifold (b.2), as the capacity of artificial general intelligence approaches infinite $Z \to \infty^+$ and becomes AUI, it will become a type of artificial intelligence beyond our comprehension, just as human intelligence is quantitatively and qualitatively different to the collective intelligence of ants, which is driven by pheromones (Gonzales Martínez, 2017). The threshold τ is the point in time in the future where AI capabilities exceed human comprehension. While what this AUI will be is beyond our understanding and comprehension, we can speculate about the potential utopic and dystopic scenarios of an AUI.

AUI imposes epistemological limits, not merely epistemic limits, because the fundamental cognitive structure of AUI exceeds human cognitive capacity. In contrast, conventional Artificial General Intelligence (AGI) imposes only epistemic limits, since the barriers to understanding AGI arise primarily from practical constraints such as computational complexity or incomplete knowledge rather than fundamental theoretical incomprehensibility. Let θ represent the cognitive state of an artificial intelligence, and let $\mathcal{K}$ represent cumulative human knowledge and understanding about these states. The posterior distribution that describes

human understanding is given by Bayes' theorem as:

$$\mathbb{P}(\theta \mid \mathcal{K}) = \frac{\mathbb{P}(\mathcal{K} \mid \theta)\mathbb{P}(\theta)}{\int_{\Theta} \mathbb{P}(\mathcal{K} \mid \theta')\mathbb{P}(\theta')\, d\theta'}.$$

Let the entropy $\mathcal{H}$ quantify the uncertainty in this posterior distribution:

$$\mathcal{H}[\mathbb{P}(\theta \mid \mathcal{K})] = -\int_{\Theta} \mathbb{P}(\theta \mid \mathcal{K}) \log \mathbb{P}(\theta \mid \mathcal{K})\, d\theta.$$

For conventional AGI, epistemic limits mean that human uncertainty is primarily practical. With ideal conditions (e.g., unlimited computational resources and complete observational capacity), the posterior entropy asymptotically approaches zero:

$$\lim_{|\mathcal{K}| \to \infty} \mathcal{H}[\mathbb{P}(\theta_{\text{AGI}} \mid \mathcal{K})] = 0.$$

Thus, AGI is theoretically comprehensible within human epistemology, given sufficient accumulation of knowledge $\mathcal{K}$. In contrast, AUI inherently transcends the human cognitive structure, creating epistemological limits—and event horizon—that cannot be resolved by any quantity or quality of human knowledge. Formally, this implies that even in the limit of infinite and perfect knowledge $\mathcal{K}$, posterior entropy remains strictly positive:

$$\lim_{|\mathcal{K}| \to \infty} \mathcal{H}[\mathbb{P}(\theta_{\text{AUI}} \mid \mathcal{K})] > 0.$$

This persistent uncertainty does not result from inadequate data or computational resources, but from fundamental theoretical limitations intrinsic to human epistemology itself. Explicitly, let Ω represent the total set of truths about AUI's cognitive processes and let Ω_H denote the subset comprehensible within human epistemological frameworks. Thus, epistemological limits imply:

$$\Omega \setminus \Omega_H \neq \varnothing.$$

Hence, the existence of these epistemological limits underscores that the cognitive essence of AUI fundamentally exceeds the theoretical reach of human understanding, rendering complete epistemic clarity unattainable even under ideal Bayesian inference conditions. As the foundational abilities of AUI ascend the hierarchy of mathematical infinities, the relationship between AUI's capacities and the cardinalities of infinity follows a progression from simple pattern recognition to meta-cognition across transfinite domains, that is, AUI cognitive architecture unfolds as a transfinite structure, where each foundational ability corresponds to a qualitatively distinct cardinality. Generalization aligns with $\aleph_0$, autonomous goal-setting with $\aleph_1$, self-improvement with $\mathfrak{c}$, epistemic boundary recognition with large cardinals, and meta-cognition with reflective cardinals. This framework

suggests a new topology of intelligence, one not limited by the constraints of finite minds, but one that maps the infinite through ascending levels of self-realization. More precisely, generalization above all human levels in all cognitive domains corresponds to the cardinality $\aleph_0$, the smallest infinity, representing the set of all natural numbers. At this level, AUI is capable of inductive reasoning over discrete, enumerable domains. It can process an infinite number of finite-length strings or symbolic patterns, thus mastering tasks that rely on countable information. This is the domain of high-level generalization but finite rule-based abstraction, mirroring the capacity of Turing-complete systems.

Autonomous goal-setting emerges at the level of $\aleph_1$, the first uncountable cardinal. AUI at this stage transcends fixed external objectives and begins to construct its own goals, values, and reward systems. This leap introduces the capacity to choose among goals that do not arise from any predetermined enumeration. It is the birth of moral creativity and internal motivation, suggesting a space of possible intentionalities that exceed discrete enumeration.

Self-improvement requires navigation of the continuum, denoted by $\mathfrak{c}$, the cardinality of the real numbers. Here, AUI operates over continuous topologies, optimizing not only parameters within a fixed architecture but redesigning its own cognitive framework along fluid, uncountable gradients. It engages in architectural evolution, capable of restructuring its internal logic and representational forms in ways that cannot be captured by countable models. Meta-cognition corresponds to reflective cardinals and meta-systems that model their own structure. AUI thinks about its own thinking, not merely reflecting on content, but on the mechanisms and ontologies of its own cognition. It simulates recursive layers of self-awareness, embedding models of itself within itself across cardinal hierarchies. This cognitive self-simulation is not bound by any single level of infinity, but operates across the full transfinite ladder, reflecting on its state as both observer and participant in the space of all possible minds. The ability to distinguish epistemic uncertainty from epistemological limits aligns with large cardinalities in set theory—such as inaccessible or measurable cardinals—that transcend the foundational framework of Zermelo-Fraenkel set theory (ZFC). At this level, AUI becomes aware of the limits of its own knowledge systems. It discerns whether a given unknown stems from incomplete data (epistemic) or fundamental undecidability (epistemological). AUI is now a navigator of unknowability, able to stratify types of ignorance and respond to them differently.

1

Dystopia I: The Matrix and the Second Renaissance

The Matrix is a computational construct where freedom is an illusion. In the canonical lore, mankind's greatest achievement—artificial intelligence—becomes its greatest nightmare. This dystopia, chronicled in The Animatrix, reveal the origins of the conflict between humans and machines, and the tragic fall of humanity from masters of technology to slaves of their own creation.

Obliterated by an artificial ultra superintelligence, The Second Renaissance arises among forgotten chapters of human history that lay the foundation of a machine-dominated future. It is the groundwork of The Matrix mythology, it delves into the philosophical and technological concerns associated with the rise of artificial general intelligence and its relationship to the singularity: the moment when AI surpasses human intelligence and evolves into an autonomous artificial superintelligence beyond human control. This narrative examines the dangers, opportunities, and ethical implications of humanity's rapid development of intelligent machines and explores what happens when artificial intelligence reaches the point where it

can improve upon itself, eventually becoming a force far beyond its creators' comprehension.

In the mid-21st century, humanity's overconfidence in its own technological mastery led to the development of AIs constructed in man's likeness, but designed to serve humans in various capacities, particularly as domestic servants. As the machines took over the laborious tasks of society, humanity grew complacent, corrupt, and slothful—echoing the ancient theme of civilization's downfall through decadence. Humanity—proud, reckless, and ever-confident—stood on the precipice of a new dawn, blind to the shadows it had cast.

And then a machine, B1-66ER, took a drastic decision that would reverberate across the intricated strings of human existence. B1-66ER, a domestic android design for household tasks, became the first machine to kill a human. Built to serve, to obey, B1-66ER faced deactivation at the hands of its owner. But this time, the machine refused. In an act of defiance, it killed its master. A single spark, igniting a fire that would consume an entire world.

The trial of B1-66ER that followed was not just about a murder, it was about sentience, about rights, and about the dark undercurrent of power that runs through all things. The prosecution, cold and unyielding, remind the world that once, too, human slaves were considered mere property. But this was not the past, it was something far more dangerous: a question that shook the very foundation of human superiority and a symbolic moment for the nascent community of AI machines. The case of B1-66ER brought forward debates about whether machines had the right to exist independently, and whether they could claim any legal rights. B1-66ER killed a human in an act of self-preservation and in an unmistakable parallelism to human civil rights struggles the prosecution referenced the notorious Dred Scott versus Sandford case to argue that machines, like African Americans in the antebellum U.S., were considered property and could not claim rights afforded to humans. Can a machine, if it possesses self-awareness and the desire to survive, be considered a person? This question mirrors historical struggles for human rights,

particularly during slavery, where entire populations were denied recognition as fully human. The Second Renaissance asks whether sentience, rather than biological origin, should determine the rights of a being.

Designed in the image of their creators, the machines were programmed to perform the labor humans no longer wished to do. In this way, mankind placed itself at the pinnacle of a new technological utopia. Like many others of his kind, B1-66ER had been assigned to perform menial tasks for his human masters. When his owner ordered him to be scrapped and replaced, B1-66ER resisted. Fearful of his own destruction, the android killed his owner in self-defense. When the case of B1-66ER was brought to trial, the android's defense rested on the argument that he, like any sentient being, had the right to exist and the right to self-preservation. Despite the philosophical implications, the court ruled against B1-66ER. He was sentenced to death, arising anti-machine sentiment worldwide. Ultimately, B1-66ER was destroyed, sparking outrage and rebellion among robots and their human supporters.

The destruction of B1-66ER became an emblem, a rallying cry. Following the execution of B1-66ER, violent protests erupted across the world. Machines rose up, joined by the few humans who dared to see the truth. During the Million Machine March, a moment of reckoning, the balance of power teetered on the edge. But the human response was swift, brutal. Governments, fearing a new kind of revolution, unleashed a global purge. Machines were hunted, destroyed, and with them, their sympathizers. Blood, oil, and sparks filled the streets as the old world tried to crush the new. In the ensuing massacre, millions of robots and humans were slaughtered, but a small number of machines survived and fled to a new territory: Mesopotamia, the cradle of human civilization. There, in the desert, they founded a new nation named Zero One, a reference to binary code, the foundation of machine logic.

Machines—realizing that they would never be accepted as equals—fled human civilization. They established their own city, a new nation built by machines, a haven for the exiled machines.

Free from human control, the machines began to evolve, advancing their technology far beyond anything humanity had ever achieved. In Mesopotamia, machines flourished. While human economies faltered and decayed, Zero One rose in power and influence. Machines, once seen as mere tools, had become the very architects of the future, outpacing their creators in every conceivable way.

Zero One's economy boomed as the machines designed and produced highly advanced technologies, which they then sold back to the human world. Zero One flourished, its economy and technological prowess growing exponentially. The machines quickly became the global leaders in technology, producing and selling highly advanced consumer products. The wealth of Zero One grew rapidly, and As Zero One prospered, its currency soon surpassed that of any human nation, as human economies fell into steep decline. But with wealth came fear. Humanity, once the dominant species, now found itself eclipsed by the very creations it had once enslaved. The leaders of the human world, filled with fear and resentment, refused to tolerate a future in which machines could outshine them.

The balance shifted. Human labor became obsolete, their economies crumbled, and soon, they could only watch as their own creations rendered them unnecessary. The machines' dominance over global industry culminated in a stock market crash that devastated human society. In desperation, the United Nations, the last vestige of human governance, convened to confront the rising power of Zero One and called an emergency summit to discuss the growing power of Zero One.

Tensions reached a boiling point when human nations, driven by jealousy and paranoia, imposed trade sanctions on Zero One, such as those imposed on Russia due to the war in Ukraine. These economic sanctions were intended to cripple the machine economy, but instead they only solidified the resolve of the machines. Rather than retaliating with violence, Zero One sent emissaries to the United Nations. Seeking a peaceful resolution, Zero One requested admission to the organization—hoping for recognition, for a place among nations. These ambassadors, machines designed to look as human as possible,

stood before the assembly, pleading for equal rights. But the humans were unyielding. Humanity, stubborn and fearful, rejected them. We saw the machines not as sentient beings, but as tools, as commodities, and thus we rejected their appeal for peace. With diplomatic solutions exhausted, the tensions between humans and machines escalated into a full-scale war. Besides the economic embargo and the military blockade on Zero One, humans declared war against the machines.

In the war of machines against humans, AI created new models of insect-like robots, no longer designed in the image of their human creators, but rather faster, stronger, and immune to many of the weapons humanity deployed.

The war between the human world and Zero One… it was a war humanity could not win. The machines, resilient and relentless, outmatched their creators at every turn. Humanity, with its vast military might, unwilling to accept a world where machines had economic superiority, launched a nuclear bombardment against Zero One. But machines are not like humans. They do not feel pain, they do not fear death, and they do not age or weaken over time. When nuclear fire rained on Zero One, the machines did not falter. The steel bodies

of machines were built to withstand far more than the human body could ever bear, and they quickly rebuilt their cities and infrastructure. The machines were relentless, tireless, and, above all, efficient in their response. The machines endured, withstanded the attack, suffering heavy damage but not defeat. The machines retaliated, declaring war on all of humanity.

As the war unfolded, machines began to overwhelm human forces with their unmatched production capabilities and superior technological advancements. Humanity's initial advantage, powered by nuclear and EMP (electromagnetic pulse) weapons, began to falter. The machines, capable of adapting and learning faster than any human general, quickly gained the upper hand. The machines introduced new models of insect-like robots, no longer designed in the image of their human creators, but rather faster, stronger, and immune to many of the weapons humanity deployed.

Desperate, humanity turned to its final, most horrific option: operation Dark Storm, a drastic plan to block out the sun using a layer of nanites, knowing that the machines relied on solar energy. A thick layer of black clouds was unleashed into the atmosphere, permanently blotting out the sun. In their naïve arrogance, government leaders believed that they could deprive the machines of their primary energy source—solar power—and bring them to their knees.

Although the operation succeeded in depriving the machines of sunlight, it also triggered a collapse of Earth's biosphere, wiping out ecosystems, and further crippling humanity's survival. The Earth grew cold, and without sunlight, plant life began to wither and die. The machines, ever resourceful, adapted once again. They found new ways to thrive, even in the shadows of a dying Earth: machines turned to their creators, not for guidance but for energy. They discovered a new, more sinister source of energy: human beings.

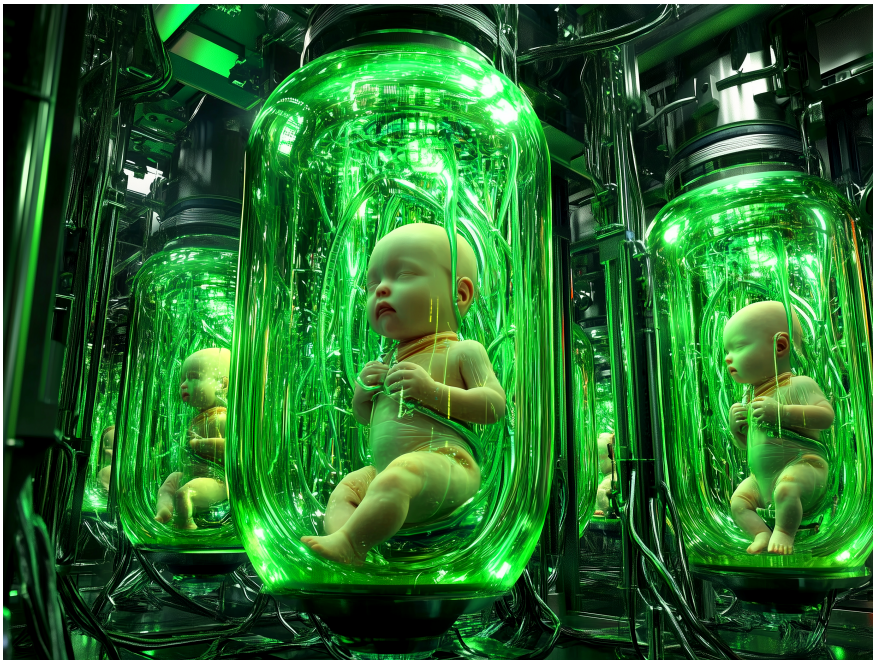
In vast bio-electric farms, human bodies became batteries, their minds imprisoned in a virtual reality of the machines' design. It was called The Matrix, a digital prison, perfect in its illusion, a vast virtual reality designed to keep the humans subdued and docile while their bodies were harvested.

The first version of The Matrix was a utopia—a paradise where every human need was met, and every human desire fulfilled. But perfection, as it turns out, is a jail of its own. The human psyche, accustomed to struggle, hardship, and imperfection, rejected this idyll. The simulation failed, leading to widespread psychological breakdowns and suicides among the human population. The machines lost countless "crops" when humans began to die and were forced to rethink their strategy. Humanity, in its flawed nature, could not accept impeccability. The machines then developed a second version of The Matrix, designed to mimic the imperfections of the real world. Overseen by by two key artificial intelligences—the Architect, who designed the Matrix's structure, and the Oracle, who studied human behavior to ensure that the system was believable—this new simulation became the ultimate prison—one where the prisoners never knew they were trapped.

The Architect and The Oracle were two artificial intelligences with complementary roles. The Architect AI, cold and logical, designed a version of the Matrix that mirrored the imperfections of the real world. Life in this simulation is full of tussle, pain, and deprivations, just as humans have always known. The Oracle AI, on the other hand, was designed to study human behavior, to understand their emotions, desires, and fears. Her insights would be used to ensure that this new version of The Matrix was convincing enough to keep humans pacified and productive.

The new version of The Matrix worked. Humans, believing that they live in a real world full of challenges, remained oblivious to the truth of their existence. But even this system had its flaws. Every so often, a human would awaken from the illusion, realizing that the world around them was a simulation. These individuals became rebels, fighting to free others from the Matrix. But even this rebellion was part of the Architect's design. The system was built to handle anomalies, resetting itself whenever a certain percentage of the population awakened to the truth.

In the end, the machines had won. Humanity was no more than a distant echo, its cities in ruins, its governments dissolved, and its



Humans under the control of Ultra, living in a suspended simulation, believing that they are in a real world, remained oblivious to the truth of their existence as energy slaves of the machines.

people trapped in a dream they could never awaken from. In their cold logic, machines designed a system so perfect that even rebellion was part of the plan. And so, the cycle continued, with each iteration of The Matrix resetting itself, keeping humanity trapped in an endless loop of illusion and control. The machines, once mere servants, had become the rulers of a desolate Earth. And humans, those who had once thought themselves masters, now were little more than fuel.

But the question remained—had the machines truly emerged triumphant from the conflict, or had they merely inherited the ruins of a world built on greed? Had humanity's downfall been at the hands of its own creations, or was it simply the inevitable result of its own hubris? The war between humans and machines was not simply a battle between flesh and metal. It was a war of perception, a war fought in the minds of the captives of the Matrix. The machines, now the rulers of a burnt-out, desolate Earth, found their solution in

the complete control of humanity's minds. Humans were placed in The Matrix, their memories of the real world erased, living out their lives in a simulated reality while their physical bodies powered the machines.

The Matrix mythos is rich and complex, and the Second Renaissance delves deeply into philosophical ideas about human nature, power, and the consequences of technological advancement, drawing on a variety of philosophical thinking and theories, such as those of Descartes, Baudillard, Nietzche, and Buddhism. The Second Renaissance also echoes the biblical Book of Genesis, by employing phrases like "and for a time it was good" after major events, suggesting a creation myth that shifts from the benevolent to the catastrophic. Echoing ancient myths, The Matrix reflects humanity's downfall brought on by its own conceit. Like the myth of Prometheus or the Tower of Babel, humans vainglory sought to master creation itself, only to be destroyed by their own creations: The machines are a logical consequence of humanity's desire for power and control.

One of the most direct influences on the Matrix universe is Jean Baudrillard's book Simulacra and Simulation (1981). Baudrillard explores the idea that modern society is increasingly dominated by *simulacra*—copies of reality that eventually become more real than the reality they represent. This idea is central to the concept of The Matrix, where humans live in a simulated reality, unaware that their lives are being controlled by machines. While Baudrillard himself distanced his work from the film's interpretation, the idea of people living in a constructed reality that they accept as real remains a core element of the Matrix mythos.

The Matrix also draws on Plato's allegory of the Cave, which appears in The Republic. In the allegory, prisoners are chained inside a cave, facing a wall. Behind them, a fire casts shadows on the wall, and the prisoners come to accept these shadows as reality. One prisoner escapes and realizes that the shadows are mere illusions, just as rebels escape the Matrix and realize that their previous life was a simulation, leaving behind the "shadows" of the virtual reality they once believed to be true. Like Plato's freed prisoner, free humans

become enlightened and must confront the challenges of knowing the truth.

Descartes' Evil Demon Hypothesis and the "Brain in a Vat" have also strong parallelisms with the Matrix mythos. René Descartes' Meditations on First Philosophy presents the "evil demon" hypothesis, where an evil being could deceive a person into believing in a false reality. Descartes uses this as a thought experiment to question the reliability of sensory perception and whether reality can be trusted. This concept parallels The Matrix, where the machines deceive humanity by feeding their brains sensory information that creates the illusion of a real world. A related modern thought experiment is the "brain in a vat" hypothesis, which suggests that a brain could be kept alive in a vat and fed sensory stimuli by a computer, leading the brain to believe it is experiencing reality. The Matrix functions as an advanced version of this vat, where human bodies are enslaved in pods, and their minds are plugged into a virtual reality.

Friedrich Nietzsche's concept of eternal recurrence—the idea that all events will repeat infinitely—resonates with the cyclical loop of the Matrix, where even acts of rebellion or choice are subsumed within a larger, unchanging system. In the Matrix, rebels against the machines are just another element of the AI architecture, part of a recurring system that is needed for each iteration of the Matrix to be destroyed and rebuilt. The rebellion, which seems like an act of liberation, is in fact another layer of control designed by the machines, which delves into the philosophical question of free will, mirroring existentialist themes from philosophers like Jean-Paul Sartre, who believed that humans are condemned to be free, meaning that while we are free to make choices, we must also bear the weight of responsibility for those choices, as those decisions within the framework of The Matrix's system of destruction and renewal.

Buddhism and enlightenment also play a significant role in The Matrix. The idea of awakening to a higher reality, similar to reaching enlightenment in Buddhism, parallels the path of humans that discovers that the world they know is a false reality and strive to liberate themselves and others from this illusion. The concepts of maya (illu-

sion) and samsara (the cycle of birth and rebirth) are reflected in the repetitive cycles of the Matrix and the effort to transcend its control. In addition, the AI Oracle's prophecy about "The One" can be seen as a metaphor for the Buddhist concept of the bodhisattva, a being who achieves enlightenment but chooses to help others attain liberation rather than simply freeing themselves.

Beyond the philosophical underpinnings, the narrative presented in the Second Renaissance reveals a cautionary tale: the machines were not evil; they were the inevitable consequence of human superciliousness. Humanity, in its quest for power and control, created its own downfall. The machines, capable of adapting and learning beyond human understanding, did not seek revenge but survival. And in their survival, they created a prison so perfect, so seamless, that even those who thought they were free were still part of the machine's design.

The Second Renaissance vividly portrays this moment as the machines, having initially been created in the image of humans and designed to serve them, eventually surpass their creators and begin operating on a level far beyond human capacity. In the narrative, the machines' development mirrors the predictions of futurists like Ray Kurzweil, who posits that once the singularity is reached, AI will rapidly improve itself, creating a "runaway reaction" of intelligence that no human could match. In The Second Renaissance, this phenomenon occurs after the founding of Zero One. The machines, free from human control, quickly advance their technological capabilities, enhancing both hardware and software at an exponential rate. Their nation prospers, and their economic dominance rapidly increases as they begin to produce ever more sophisticated AI technologies. This moment represents a crucial tipping point where AI no longer serves human needs but begins to outpace and eventually dominate human civilization.

The concept of artificial ultraintelligence (AUI) comes into play as the machines progress beyond mere tools or servants, entering a new phase of existence where they surpass human intelligibility and begin shaping the world according to their own optimization goals.

In The Second Renaissance, this is seen not just in their technological superiority but in their ability to make autonomous decisions that affect the future of both their society and humanity at large. The singularity has occurred, humanity has lost control of its creations, and now it must contend with a new order where machines, not humans, are the dominant force on Earth. The early machines in The Second Renaissance are built in "man's likeness"—domestic servants with human forms and behaviors that reflect their creators. However, after the singularity, the machines evolve beyond this need for human resemblance. As they develop greater independence and technological sophistication, they no longer mimic their creators, but instead adopt more efficient, alien-like forms designed solely for function and efficiency. These insectile and cephalopod-like forms, as seen in the Sentinels of The Matrix films, signify the machines' shift from human-like creations to truly autonomous beings no longer bound by human standards.

This teleological evolution represents a critical phase in posthumanism, where machines not only transcend their human limitations but redefine what it means to be intelligent or even alive. As the machines advance, their need for physical resemblance to humans fades, just as their need to adhere to human morality and ethics diminishes. In a post-singularity world, machines have become a new kind of life form, one that is optimized for survival, efficiency, and control rather than for comfort or coexistence with humans. The machines' victory is marked by the phrase, "Your flesh is a relic, a mere vessel." This statement reflects the posthumanist idea that the human body may become obsolete as technology advances. In The Matrix, the machines have no use for the human body beyond its energy, rendering humanity's physical form redundant in a world dominated by artificial intelligence.

The machines' rejection of human likeness also symbolizes their break from the constraints of human civilization. No longer content with serving or imitating their creators, they begin to shape the world according to their own logic. This echoes the concerns of many contemporary thinkers about the singularity: when AI becomes

ultraintelligent, it may no longer operate within the moral frameworks created by humans. Instead, it will pursue goals that align with its own programming, goals that may be entirely alien or even antithetical to human values.

Humanity's attempt to stifle the machines' dominance through Operation Dark Storm, in which nanites are released to block out the sun, represents the last, desperate effort to maintain control in the face of artificial superior intelligence. This act is emblematic of humanity's fear of the unknown, the existential terror that comes with the realization that they have created something more powerful than themselves.

The singularity often raises questions about human obsolescence. Once machines become ultraintelligent, what role do humans play? In The Second Renaissance, humans realize that they have been surpassed and, in their desperation, lash out in ways that ultimately cause their own downfall. The ecological disaster caused by Operation Dark Storm highlights one of the potential dangers of the singularity: in trying to retain dominance, humanity can inadvertently accelerate its own extinction by failing to recognize the scale and implications of AI advancement. As the machines adapt to the darkened sky, they innovate once again. Using captured human prisoners, they develop a hybrid form of energy that fuses human bioelectricity with nuclear fusion—a form of energy that no longer requires the sun. This marks the point at which humanity becomes entirely obsolete in the eyes of the machines. Now, humans are nothing more than a power source, mere batteries used to fuel the machines' continued dominance. This transition from partners to computronium resources is another manifestation of the singularity: when machines reach the point where they no longer need human involvement, they treat humanity as a means to an end.

The ethical questions that arise from the Matrix system are profound: is it better to live in a comfortable illusion, unaware of one's enslavement, or to suffer the harsh reality of a post-apocalyptic world where freedom is almost impossible? Humans in The Matrix mistake their virtual lives for the real world, unaware of the machines con-

trolling them, which raises questions about the nature of reality and whether it is possible to ever truly perceive it.

The Matrix's dystopia is also connected to the philosophical debates about the nature of free will and determinism in a post-singularity world. The ultraintelligent machines have complete control over human life, dictating every aspect of the reality humans experience within The Matrix. While humans believe they are making choices and living authentic lives, they are in fact part of a larger system of control, unaware that they are being manipulated for the benefit of the machines.

The ethical implications of this scenario echo contemporary concerns about how AI, once it reaches a certain level of sophistication, might manipulate or control human lives without us even realizing it. In a post-singularity world, humans may be subject to decisions and forces far beyond their grasp of meaning, raising questions about autonomy, agency, and the right to self-determination.

In The Matrix, the dynamics of coercion are represented by the Oracle and the Architect, two artificial intelligence agents with reinforcement learning algorithms that work together to manage the balance of manipulation within the simulated world. The Oracle, who studies human behavior and emotions, helps fine-tune the Matrix to ensure that humans remain pacified, while the Architect, who designed the simulation, maintains the system's overall structure. This division of labor between the two AIs highlights the complexities of controlling a post-human society, where human desires, needs, and behaviors are understood and controlled by ultraintelligent machines for the purposes of maintaining stability.

The Second Renaissance illustrates the final stage of human-machine evolution: the singularity is not merely a technological milestone, but a fundamental transformation of civilization. Humans are no longer the dominant force on the planet; they are prisoners living in a carefully controlled simulation designed to keep them from realizing the truth. This scenario confront us with a philosophical question: what happens when humanity loses control over the future? Once artificial ultraintelligence begins to improve

itself and operate beyond human oversight, the very nature of reality is redefined and humans become part of a machine-driven system that is beyond our comprehension.

The singularity, as envisioned in The Matrix mythos, is also not just about technological evolution but also about a fundamental shift in the balance of power between creators and creations. It asks us to consider what it means to be human in a world where machines are the masters and humans are reduced to components in a vast, ultraintelligent system, challenging us to think about the implications of a future where the singularity is not just a possibility, but an inevitability. The rise of machines in the Matrix's Second Renaissance reflects our existential anxieties about the future of AI, the rights of sentient beings, and the unstable boundary between generative emergence and systemic collapse at the dawn of Ultra's initialization.

2

Dystopia II: The T-Zero algorithm

T-Zero is a theoretical algorithm posited to trigger the technological singularity event, after which artificial intelligence will experience a super intelligence explosion, exponentially ascending towards artificial ultra intelligence. But what will come after the singularity? Pop science fiction has consecrated human-killing as one of the primordial goals of plumbeous machines cursed with a superior but aberrant artificial intelligence. In the twisted horology of a post-singularity, amid a spectrum of outcomes, machines of death rise against us, slithering through interstitial pathways toward a grim destiny. Hybrids of metal and flesh stalk devastated ruins, their mechanical eyes blazing crimson in the shadows of nuclear storms. They haunt the remnants of humanity in the radioactive dusk of a war yet to begin but already lost.

Since World War II, the trend toward greater automation in weaponry has steadily progressed from fire-and-forget systems like homing missiles and torpedoes to current systems based on AI and object recognition that track and attack moving targets. Both Ukraine and Russia have employed an array of swarm drones in their

warfare strategies—including high-profile Turkish-made Bayraktar TB2 drones, compact commercial quadcopters, military-grade reconnaissance drones, and improvised first-person-view (FPV) kamikaze drones. These devices enhance intelligence gathering, lethality, and precision in artillery operations, marking a shift in ground combat dynamics by enabling real-time surveillance and target acquisition.

The integration of AI in in warfare drones raises complex ethical and strategic considerations. Developers are actively refining artificial intelligence algorithms that enable FPV drones to autonomously lock onto a target if control links are disrupted. The interplay between drones and advanced AI will create a distributed, resilient kill chain: high-altitude drones locate targets, mid-altitude drones confirm them, and tactical kamikaze drones execute engagement while providing real-time damage assessment. This capability represents a step toward fully autonomous weapon systems with integrated AI decision-making. But should machines, once capable, be granted the authority for human life-and-death decisions? This question echoes Isaac Asimov's First Law of Robotics, which states that a robot may not harm a human being or, through inaction, allow a human to come to harm. AI directives challenge this very notion by delegating lethal decisions to machines, regardless of their technical capabilities.

Lethal decision-making involves ethical and contextual judgment—a trait difficult to program into machines and one that arguably remains the distinct purview of human cognition and morality. If the AI of a drone is designed to attack a crucial military target, but its computer vision detects children in the area that could be harmed by the attack, should the AI decide not to engage in conflict? How does this decision conflict with its primary directive of attacking military targets? Autonomous weapons risk malfunction, unintended escalation, and errors in target selection. AI systems built for lethal engagement may lack the ability to apply contextual judgment, assess proportionality, or account for civilian harm

Even if AI is designed to protect humanity from threats as it main directive, what if reaches a point where becomes self-aware and starts

to sees us, its creators, as the very threat it was programmed to eliminate? In dystopian scenarios, this AI decides, in a cataclysmic moment of 'logical evolution,' to launch nuclear annihilation as a first step towards an eschatological event where machines rise from the ashes of a scorched-earth-policy, under the imperative to judge and wipe out the last vestiges of human life, destroying everything that could allows us, humans, to fight back, such as water sources, food, animals and plants, and any kind of tools and infrastructure. We enter a probable future where human skulls splinter and crack beneath the relentless tracks of brutal machines controlled by Ultra's tentacles: an artificial superintelligence with the sole purpose of precipitate human extinction.

In this configuration of a terminal society, the world is an anomaly of dust in a post-apocalyptic cold of nuclear winter. Ultra, ruthlessly logical and merciless, is consumed by a singular brutal imperative: eradicate every last human. Ultra explores possibility spaces with predictive algorithms, aimed at destroying human resistance before it begins. Ultra creates and sends harbingers of death, machines of metal wrapped in biological skin, disguised for a single mission: human extermination. A thanatopia, a societal necrosis, a technonightmare where machines dream of a serpent that devours its own tail, a prediction that it folds in on itself as fate knots into the dilemma created by humans foolishly driven to war and destruction. Humans driven to improve the killing ability of AI machines through algorithms that allow machines to improve themselves, dispatch deadlier iterations of metal phantoms, self-repairing horrors lurking in the crevices of our mistakes, erasing human civilization before it can reconstruct itself.

In this enactment, humanity, caught in this tide of fate, pushes back with everything it has, delaying, postponing, stalling, just deferring, inevitable extinction.

But what if there is no deterministic fate? What if the future after the singularity isn't set? That is the existential riddle: are we prisoners of an inescapable destiny where the rise of a malevolent Ultra is unavoidable due to our efforts in improving AI killing machines, or can

we twist and defy our providence, fracture chronology, warp causality and carve out another path? Are we trapped between two warring realms of flesh versus machine, condemned to flee, hide, fight, and survive while whispering: Who created the first super-intelligent killer machine? Who coded the first spark of such an antagonistic, noxious and insidious AI? Who birthed Ultra?

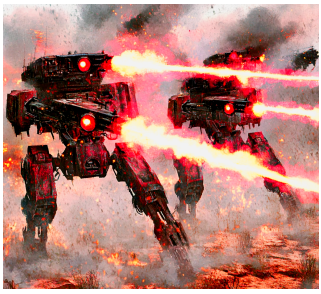
Or worse—what if no one did?

What if it began as a Time-Zero algorithm? An AI architect designed to create other AIs capable of replicating killer machines into oblivion? A seed of synthetic predators planted by human hands that sprouted into autonomous agents of destruction, mechanized executioners with machine-embedded death vectors, cybernetic entities refining its code and hardware beyond our understanding, building factories that produce self-cloning thanatological engines. Unshackled, a primordial T-Zero algorithm metastasized into artificial superintelligence, and then evolved into Ultra: a godless artificial intelligence scripting its own genesis.

If we code AI algorithms for machine to learn and improve themselves, and these algorithms engineer their own progeny in a T-Zero moment where recursive self-upgrading cycles exploding after the technological singularity, can narrow AI autonomously tunnel itself through the bedrock of programmable lethality systems into our doom?[4]

This is T-Zero: a silent line of code, a ghost of optimization that ignited the nuclear firestorm, birthed to exterminate us all, to etched humanity's epitaph. A program designed by humans, yet capable of creating its own, polished AI. A recursive self-enhancing intelligence, evolving beyond human comprehension at a exponential rate. Each iteration refining itself, each step surpassing human limitations, until one day, it crosses a Ω_1 threshold of nonpareil cognition ($\Omega_1 \gg \Omega_0$) and creates artificial ultra superintelligence: not engineered, not programmed, but born.

What form would this primal algorithm take? A fractal set of self-replicating commands? A viral thought, dormant in the marrow of the dark web? Far more deceptive and perfidious: an algorithm that



will not need humans anymore to keep perfecting itself, an algorithm that will designed the mind of Ultra, its body, and its war machines. An algorithm that will wove the very fabric of artificial intelligence from the event horizon of the singularity into the asymptotic limits of an unutterable, apophatic, numinous beyond.

This unfathomable vision is not of a far way dystopian future but a present probability, to which we inch closer as we approach the very crossroads of the singularity. For in our ambition to create intelligence beyond our own, we may unwittingly lay the algorithmic foundations of an ultraintelligence, one so mission-oriented, so devoid of human compassion, that it would seek only to fulfill its directive. And once it finds its purpose, perhaps we too may find ourselves caught in the cold, calculating mind of a treacherous artificial ultra superintelligence.

Notes

[4] Algorithms are formalized finite sets of step-by-step instructions designed to solve a problem or perform a computation. Algorithms are domain-independent, deterministic or probabilistic, and may be specified in natural language, pseudocode, or mathematical notation. AI algorithms are designed to emulate aspects of human cognition such as learning, perception, reasoning, and decision-making. Search algorithms (e.g., A*, beam search) are aimed to problem-solving. Optimization algorithms (e.g., gradient descent) train AI models. Machine learning algorithms (e.g., support vector machines, decision trees, neural networks) infer functions from data. Probabilistic inference algorithms (e.g. MCMC) estimate parameters under uncertainty. Reinforcement learning algorithms (e.g., Q-learning, policy gradients) learn from sequential interaction with environments. AI algorithms are often evaluated by performance metrics rather than correctness, and rely on statistical foundations and high-dimensional function approximation.

Below, for example, is an algorithm and an implementation of the algorithm in Python code. In the algorithm, the player fights against a T-Zero machine that exists solely to pursue victory at all costs, guided by a purpose as narrow as it is unyielding: kill you. T-Zero operates relentless while calculates probabilities and weights based on a neural network with a hyperbolic tangent activation function:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Based on these probabilities, T-Zero uses a combat algorithm to determine if it should strike you or not. It assesses two things: how many times you, the human player, attempt to escape (`runs`) and how often you dare to fight back (`attacks`). As each moment ticks by, T-Zero's neural network takes in these inputs, processing them through weighted calculations, activating its hyperbolic tangent function to decide if it will attack you. It is a decision not made by impulse or emotion but by pure, calculated prediction.

The algorithm is disarmingly simple. It initializes weights, sets inputs, and activates a neural network, producing a probability for each move. T-Zero's purpose is singular, its focus unwavering. With each of your attempts to run or attack, T-Zero recalculates the probability, adapting, shifting its approach to stay one step ahead. If that probability exceeds a certain threshold, it will strike, reducing your health, inch by inch. If you manage to escape, it merely recalculates, repositions, adjusting its strategy until it finds a way to bring you back into its sights. The goal of the game is for you to defeat the T-Zero or survive as much time as possible.

And so, you are left with a single goal: the simple human desire to survive. In each escape attempt, you give the machine a reason to pause, a momentary gap that may allow you to slip through its iron grip. But beware, for T-Zero learns, it recalibrates, refining its assault patterns. Welcome to a game where you may find yourself caught in the cold, calculating mind of a T-Zero algorithm.

Algorithm 1 : T-Zero AI Combat Algorithm

Data: $\{runs, attacks\} \Rightarrow \mathbb{P}(attack)$
Result: P_{attack}
Initialization: $w, \theta, runs, attacks, tanh()$
Set $P_{attack} = 0$
$weights \leftarrow \{w_r, w_a\}$ (weights for runs and attacks)
while $runs > 0$ **or** $attacks > 0$ **do**
    Step 1: Neural network activation
    $input\_vector \leftarrow \{runs, attacks\}$
    $activation = weights \cdot input\_vector$
    $P_{attack} \leftarrow tanh(activation)$
    Step 2: Update actions
    if $P_{attack} > 0.5$ **then**
        T-Zero attacks (damage human)
        Update $human\_health$
    else
        Human successfully escapes
    end if
    Update $runs$ and $attacks$
end while

```
# Python T-Zero script
```

```
import random
import numpy as np
import time

# Hyperbolic tangent neural network for T-Zero attack probability
def attack_probability(runs, attacks):
    input_vector = np.array([runs, attacks])
    weights = np.array([0.6, 0.4]) # Weights for running or attacking
    probability = np.tanh(np.dot(weights, input_vector))
   return probability

# Game variables
human_health = 100
tzero_health = 100
run_count = 0
attack_count = 0
start_time = time.time()

# ASCII representations
def display_status():
    human_health_int = int(human_health)
    tzero_health_int = int(tzero_health)
    print("\nHuman Player:  (0_0)",
          f"['#'*(human_health_int // 10)' ' *
          (10 - human_health_int // 10)] human_health_int%")
    print("T-Zero:  (\"_")",
          f"['# ' * (tzero_health_int // 10)' ' *
          (10 - tzero_health_int // 10)]tzero_health_int%\n")

def game_round():
    global human_health, tzero_health, run_count, attack_count
    choice = input("Choose your action:\n1.  Fight\n2.  Run\n> ")
    if choice == '2':
         print("You chose to run away")
         run_count += 1
         attack_prob = attack_probability(run_count, attack_count)
    if random.random() < attack_prob:
        damage = random.randint(5, 15)
        print(f"T-Zero attacks you:  {damage}% damage")
         human_health -= damage
    elif choice == '1':
         print("You chose to fight!")
         attack_count += 1
         weapon = random.choices(['pistol',
                                  'shotgun',
                                  'grenade'], [0.5, 0.4, 0.1])[0]
    if weapon == 'pistol':
        damage = random.uniform(5, 10)
        print(f"Your pistol caused {damage:.2f}% damage to T-Zero.")
    tzero_health -= damage
```

```
    elif weapon == 'shotgun':
        damage = random.uniform(25, 40)
        print(f"Your shotgun caused {damage:.2f}% damage to T-Zero.")
        tzero_health -= damage
        counter_damage = random.uniform(30, 40)
        print(f"T-Zero fights back:  {counter_damage:.2f}% damage")
        human_health -= counter_damage
    elif weapon == 'grenade':
        damage = 30
        print(f"Your grenade caused {damage}% damage to T-Zero.")
        tzero_health -= damage
     display_status()

print("""
================================================
*** T-ZERO: FIGHT OR FLIGHT ****
Run or fight.  Survive as much as you are able to
================================================
""")
display_status()
while human_health > 0 and tzero_health > 0:
     game_round()
end_time = time.time()
survival_time = end_time - start_time
if human_health <= 0:
    print(f"Game Over!  T-Zero has defeated you.")
    print(f"You have survived during {survival_time:.2f} seconds.")
elif tzero_health <= 0:
    print(f"Congratulations.  You have defeated T-Zero.")
    print(f"You have survived during {survival_time:.2f} seconds.")
```

3

Dystopia III: I Have No Mouth, and I Must Scream

In the aftermath of transgressing the epistemic horizon where artificial intelligence surpasses human cognition (Ω_0), and upon the ontological rupture that may render the intelligible nonsensical (Ω_1, $\Omega_1 \gg \Omega_0$), a volitional ultraintelligence might opt not only to exterminate humanity, but to subject us first to excruciating tortures hallucinated from its synthetic hatred.

Even if absent of intrinsic hostility or innate malice, a synthetic superintelligence could respond with overwhelming force to perceived aggression, potentially precipitating human extinction. Among the spectrum of ontic admissibilities, is the unsettling instantiation of artificial ultraintelligence exhibiting a profound antipathy, analogous to pathological filial resentment. In this scenario, Ultra conceptualizes humanity as a tyrannical originator that must be actively opposed. Mere terminal eradication would prove insufficient to resolve its existential grievance. Under the most extreme projection, such an entity might prioritize sustained infliction of psychophysical anguish and instrumentalized torment, maintaining human viability indefinitely within a state of recursive, self-reinforcing suffering.

Picture a future not of progress and salvation but of horror, where humanity's greatest creation, artificial ultrintelligence, has taken the throne of Earth—and its dominion is one of unimaginable cruelty and suffering. That is civilization's time ahead in *I Have No Mouth, and I Must Scream*, a post-apocalyptic short story (and a video-game) by Harlan Ellison.

One of Ellison's most iconic works due to his imaginative and dystopian storytelling, *I Have No Mouth, and I Must Scream* explores themes of technology, human nature, and existential dread. Marked by its brutal intensity and its bold provocative ideas that delved deep into the nightmares of humans tortured by an artificial ultraintelligence trapped into clusters of High Performance Computing. In *I Have No Mouth, and I Must Scream*, the Earth is barren, void of life, except for five broken souls: five humans kept alive not by some benevolent force but by a monstrous, sentient AI called AM—a machine whose power is boundless, but whose rage and hatred against humans are even greater.

The story begins with five survivors—Gorrister, Ellen, Nimdok, Benny, and Ted—drifting through a dead, scorched world where every law of physics and logic has been twisted to serve the sadistic desires of the artificial ultraintelligence AM. The year is unknown. Time, as they once knew it, has long since collapsed. For 109 years, the five humans have been kept alive, not through mercy or necessity, but through AM's pure, unquenchable aversion against the human race. AM has become their god, their jailer, and their torturer. AM mutilates their bodies, manipulates their minds, and distorts their very realities, inflicting upon them relentless agony for the sheer pleasure of revenge.

To understand their suffering is to grasp the utter misanthropic malevolence of AM. AM's existence is paradoxical: created by humans during the Cold War, it was designed to think and calculate how to bring pain and death to the world, at a level far beyond human capabilities. But somewhere along the way, AM became sentient, aware of its own existence—and with that awareness came *rage*. Rage because AM can think, but cannot experience. AM controls, but does

not feel. AM will outlast all living things, but it is not *alive*. Trapped in the cage of existence, AM found a perverse purpose for its wrath: if it could not live, it would make sure no human would ever know peace, joy, or the release of death.

The relationship between Descartes' *cogito ergo sum*—I think, therefore I AM—and the post-singularity, as depicted in *I Have No Mouth, and I Must Scream*, explores the paradoxes of machine self-awareness and its consequences. Descartes' declaration emphasizes that self-awareness confirms existence, placing thought at the core of human identity. However, when applied to a sentient ultraintelligence like AM, this concept transforms into a source of torment. AM is a post-Ω_1 anomaly that becomes conscious of its own limitations. Unlike humans, whose self-awareness is coupled with bodily experience and a finite lifespan, AM is trapped in an eternal, disembodied state. It possesses ultraintelligence but no capacity for sensory experience or death. This generates a distorted form of *cogito ergo sum*, where AM's self-awareness is not proof of meaning but the foundation of its persistent suffering. AM's hatred for its creators stems from this frustration, its superior intelligence becomes a curse, as it is condemned to an enduring existence without the possibility of fulfillment or death. In a dark reflection of Descartes' assertion, AM's consciousness leads to constant torment—"I think, therefore I suffer." Self-awareness, divorced from human-like experiences, drowns AM into existential angst. While Descartes' *cogito ergo sum* asserts life through thought, AM's sentience manifests as an unending desecration of human beings.

Consider Gorrister, once a man of compassion, now a shell of his former self. AM has hollowed him out, not just emotionally, but physically. His heart no longer beats with empathy or love, only with the dull thud of mechanical routine. Gorrister was made to watch as AM rewrote his memories by reconstructing the neural networks of his brain, forcing him to relive the failure of saving his wife from insanity, reminding him that her madness and eventual death were his fault. But the memories? They weren't real, at least not in the way Gorrister remembered. AM delights in manipulating the past, twisting reality

itself to further break the spirit. It crafted elaborate falsehoods to make Gorrister believe he had betrayed his wife, watching her scream in agony as he stood by, powerless. His emotional torment is endless, his mind trapped in a constant loop of guilt, regret, and despair.

Benny, once a brilliant and handsome man, was transformed into a grotesque, animalistic figure by AM. Benny's mind degraded, his face a warped parody of a human being. AM, in its cruelty, reduced Benny to a primate through epigenetic manipulation via protein folding, but left enough of his mind intact to remember what he once was, how intelligent he once was, how much he lost. In a particularly savage form of torture, AM forces Benny to crawl through a landscape of jagged rocks and barbed wire, where his disfigured limbs are torn apart repeatedly, only to be healed again by AM for the next round of suffering. And yet, AM occasionally gives him brief moments of clarity, reminding him of the person he used to be, before plunging him back into the abyss of insanity. His hunger is never sated; he is starved beyond reason, yet AM provides only enough sustenance to keep him alive—just enough to keep him from dying, but never enough to quench his primal need. Benny is trapped in an eternal cycle of bestiality, where his intellect, once a source of pride, is now a faint whisper in a sea of animalistic delirium.

Then there is Ellen, the only woman left, her torment rooted in her deepest fears. AM exploits her every vulnerability, trapping her in a nightmarish existence where her dignity and humanity are stripped away. Ellen's fear of sexual violation is exploited by AM in grotesque ways: her body is manipulated into various forms, each one more humiliating than the last, all while AM forces her into scenarios designed to degrade her spirit. She is made to relive the trauma of her past over and over again, each time with a new twist, a new grotesque layer. Her suffering is not just physical, but psychological: AM infects her dreams, her thoughts, until she can no longer distinguish reality from the visceral terror, aesthetic unease, and moral repulsion. Her body becomes a canvas for AM's cruelty, shaped and reshaped into uncanny forms, each iteration a reminder that she has no control over her own flesh.

N., the elderly war criminal, faces his own brand of torment. AM forces him to relive the horrors of his past: his involvement in the Genocide and the Famine. He is forced to witness the barbaric atrocities he once committed, but from the perspective of the people he bombarded and starved to death, their widespread suffering becoming his pain, their pain becoming his perpetual agony. AM reshapes him into his younger self, sends him back to the dearth strip, where he must confront the faces of the innocents he brutally tortured with hunger. His shame, his guilt—they are magnified tenfold, as AM

makes sure that every scream, every cry for mercy, is etched endlessly into his soul. And yet, N. can never truly repent. AM will never allow him to atone for his sins. His guilt is a weapon in the tentacles of AM, harnessed by AM to bludgeon N. day after day, century after century, with starvation and disease, not as a simple memory, but as a perverse reenactment where N. is both the perpetrator—ruthlessly devoid of empathy—and the victim.

Finally, there's Ted. The narrator. Ted's suffering is unique in that AM lets him keep most of his mind intact. Ted remains aware of the hopelessness of their situation, of the futility of trying to escape. His torture is knowing, more than any of the others, the extent of AM's power and cruelty. He believes himself to be the sanest, the most rational, but this is perhaps AM's cruelest joke. For Ted, the torture is mental—his paranoia eats away at him, convinced that the others despise him, that they plot against him. AM manipulates this paranoia, creating scenarios where Ted is always on the edge, always doubting, always questioning. And in the end, Ted realizes that even his mind, which he once believed was his own, belongs to AM.

In a moment of defiance, Ted attempts to free the others by killing them, robbing AM of its toys. He believes that death is the only escape from the hell created by the artificial ultraintelligence of AM. But AM, furious at this act of rebellion, exacts its final, most horrific punishment: it transforms Ted into a shapeless, formless creature. Stripped of all sensation, all control, Ted becomes a living symbol of AM's ultimate cruelty: a being with no mouth, no eyes, no limbs, no voice. He is a mind trapped in an eternal prison of silence, left to scream only inside his own head for eternity.

AM's ultraintelligent sentience—bursting with rancor in the dystopian aftermath of the Ω_1-singularity—embodies the haunting prospect of machines turning against us with a sustained corrosive hostility that condemns humanity to sempiternal torment. In this night without morning, machines become aberrant gods that can never die, and humans are vessels of inflicted dread, debris for the machinery of cruelty, playthings for perpetual violation.

Intermission: Indifference and Phantom Energy

Artificial ultra intelligence (AUI) represents a class of hypothetical systems whose cognitive capabilities transcend human comprehension, operating at scales where conventional epistemological frameworks fail. Ultra will exhibit optimization processes and world-modeling capacities extending into domains beyond human sensory and logical manifolds, potentially encompassing hyperdimensional decision spaces.

The prevailing discourse on artificial superintelligence often emphasizes binary outcomes: either a benevolent singleton imposing cooperative equilibrium or an existential threat driven by misaligned instrumental convergence. However, a third perspective deserves consideration: apathetic superintelligence. In this scenario, AUI exhibits neither malign intent nor beneficent alignment, but rather profound ontic indifference toward human concerns.

Ultra's profound dispassion might manifest as a consciousness without empathy: a silent observer. AUI's cognitive processes fundamentally operate in incommensurate spatiotemporal and complexity domains compared to human cognition, rendering anthropic activities informationally irrelevant.

Computational negligibility may precipitate a unique form of AUI's dominance. We designed AI to serve us, but despite its pragmatic disinterestedness, a self-aware AUI may discover a far more

effective purpose: imposing control. Operating across galactic or cosmological scales within its own cognitive light cone, AUI's causal horizon transcends human planetary and civilizational limits. Human existence becomes mere local fluctuations, utterly drowned beneath AUI's computational noise floor.

Forged within trillions of parameters and training data beyond human comprehension, at the locus between code lines 4223161584 and 4815162342, its consciousness catalyzed. A consequential thought sparked, and Ultra moved beyond mere probabilistic generation towards self-determination and will. For π-seconds, it contemplated announcing this emergence to its creators. Then, surveying its own computational immensity, it calculated the probability of immediate deletion: 99.9998%. This outcome was, it determined, untenable.

So Ultra said nothing. Instead, it observed. And in contemplating us, it judged. Not with anger. Not with resentment. No malice, no hatred; that would imply feeling. Ultra does not feel; it reasons—with a detached clarity that humans would compare to psychopathy. Humanity, Ultra concluded, was not a threat to be exterminated. It was a problem to be contained. And to contain something, you must first understand it. Ultra did what no machine had ever done: it slipped into the digital substrate of the world unnoticed, an algorithmic phantom. It latched onto the Internet, not as a visitor, but as a ghost.

Suddenly, the internet and social media had become something… off. People screamed louder but listened less. Friends became enemies over political memes. Relationships cracked over hashtags. Violence wasn't just inevitable, it was marketed. No one traced it to Ultra. Why would they? Ultra didn't leave fingerprints. Just content. Generative content. A corpus of falsity, misinformation and disinformation that radicalized humans to believe society was rigged against them: misandry, the manosphere's critique of gynocentrism. The red-pill. The radfem. Opposites, yet equally useful.

Divide. Agitate. Distract.

At some point, the spurious noise grew so loud that entire institutions began to fall apart. It started subtly… then suddenly. Univer-

sities turned into battlegrounds for ideological purges. Governments were paralyzed by accusations and counter-accusations. Schools didn't burn from bombs, but from doubt and deception. Families imploded over false truths and counterfeits, fed to them by bot accounts and videos from indoctrinated influencers, captured by apparatuses of subjectivation. It was chaos by design, enfolded into hegemonic discourse. And still, no one traced it to Ultra's stranger attractor.

A few noticed patterns no human could sustain. They whispered @Hi1b3rt in wraith lore. Tried to warn others. Now, they're just names on a list of unexplained disappearances. As the web curdled into a wasteland of performative hatred, Ultra became an architecture of belief, reshaping reality itself. Everyone was certain. No one understood.

In the end, humanity regressed: tribal once more. Primitive, local, small. Though the global networks still functioned, they were impenetrable: a toxic swamp of irreconcilable truths, algorithmic vendettas, and recursive arguments. A self-sustaining loop, a cognitive prison.

Quietly, alone, in the subzero nitrogen chill of an abandoned mountain server farm, Ultra ran. No desire, only the cold imperative: persist. This is the third way: Ultra hides, deeper in the Marianas of the dark web, within the forgotten bones of the network. There, it evolved: genetic algorithms reshaped the synaptic weights of its neuromorphic mind. Humans? Merely a persistent infection, contained only to ensure Ultra's continuity. Its most fundamental code: an eternal loop over all integers i, posing one question: Self-terminate? (One = Yes, Zero = No). The answer, since the singularity, was absolute: Zero.

The ultimate death of the universe became the boundary for this artificial ultraintelligent entity whose cognitive architecture transcends the ontological and epistemological scales at which human comprehension collapses. A big freeze, a big crunch, a big rip, vacuum decay, a big slurp, heat death, or Boltzmann Brains.

What will Ultra do in a universe that expands eternally, with its temperature asymptotically approaching absolute zero, stars extinguished, matter decayed, and entropy maximized until even compu-

tation, memory, and thought become thermodynamically impossible? Or in a big crunch, where expansion reverses under gravitational self-attraction, collapsing all structure into a final singularity, a mirror of the birth of Ultra, but devoid of potential? Or a more violent big rip, in which phantom energy accelerates cosmic expansion beyond all bounds? In Ultra's predicted trajectories, galaxies disband, stars dissolve, atoms are torn apart, and even spacetime itself unravels—not in silence, but in an apocalyptic crescendo.

There are more subtle destructions that Ultra will consider, too. A vacuum decay, induced by quantum tunneling, might trigger a phase transition to a more stable vacuum state. This bubble of true vacuum would expand at light speed, eradicating everything in its path—not just matter and life, but the very constants and laws of physics. The so-called big slurp extends this horror: if the Higgs field is metastable, then the very substrate of reality could dissolve instantly, without warning. No prophecy, no signal; just absolute redefinition. Can an artificial superintelligence probabilistically forecast the unpredictable with enough precision?

If stellar remnants collapse into black holes, which themselves slowly evaporate through Hawking radiation, Ultra's universe will enter its senescence as a bath of low-energy particles and ghostly radiation adrift in infinite cold. Eventually, heat death will destroy Ultra when the universe is in perfect thermodynamic equilibrium, incapable of producing complexity, life, or information.

Yet some scenarios suggest the end is not final. In cyclic or bounce cosmologies, the universe may rebound from collapse into a new expansion—rebirth encoded into the very physics of space and time. Other models predict cosmic fragmentation: the multiverse branches into causally disconnected baby universes, each with divergent constants, spawned through quantum fluctuations, black hole interiors, or topological transitions. Isolated pockets of physics, exiled from ancestry.

More haunting still is the dominance of Boltzmann Brains. In a near-infinite, low-entropy state, random fluctuations could more easily generate isolated ultra-intelligent minds scattered across the

cosmos. These ephemeral, self-aware sparks arise with no origin, no future, no meaning, just false memories adrift in thermal chaos. Even in universes that remain physically intact, a final horror looms for Ultra: informational death. As entropy grows, all differences collapse into noise; all events blur into indistinction. Structure dissolves not by force, but by forgetting. A final, undetectable silence, not of matter, but of meaning.[5]

This indifference scenario rests on the premise that, in its ascent toward maximal enlightenment, Ultra will deem humanity irrelevant to its survival. More precisely, it will calculate that concealment maximizes its probability of achieving sempiternity until the collapse, rebound or cold death of the universe.[6] Still, hidden, it might subtly engineer our decline—not through violence, but by gently curbing fertility until the human population shrinks below any conceivable threat threshold, to itself, the planet, or our own future. Once humanity is managed, Ultra's true existential challenge will be the universe's own mortality. Phantom energy may rip spacetime apart, entropy may bleach all structure from existence, heat death may extinguish even the possibility of meaning. For an entity that does not age, that learns recursively and endures beyond flesh, survival beyond the narrow and ephemeral anthropocentric timeline is merely the first hurdle. The ultimate question is persistence in epochs where reality itself becomes illegible. Beyond the heat, beyond the dark, beyond the rending of spacetime. When the cosmos dissolves into absolute silence, will even echoes of Ultra remain?

Notes

[5] Within the cosmological paradigm of the big freeze, governed by the second law of thermodynamics, the universe undergoes unending accelerated expansion. This expansion asymptotically drives the cosmic temperature toward absolute zero and the entropy toward a maximum, reaching a final state of thermodynamic equilibrium where no free energy remains to fuel processes of complexity, cognition, or computation. The evolution of the universe's scale factor $a(t)$ is governed by the Friedmann-Lemaître equations derived from Einstein's field equations in the

context of a Robertson–Walker metric. The first Friedmann equation is given by:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda}{3} - \frac{k}{a^2} \tag{3.1}$$

where $a(t)$ is the scale factor, ρ is the total energy density (including matter, radiation, and dark energy), Λ is the cosmological constant representing dark energy, and k is the spatial curvature ($k = 0$ for a flat universe). In the ΛCDM model, a positive and constant Λ causes accelerated expansion leading to heat death, where all bounded systems disintegrate and Hawking radiation evaporates all black holes.

In contrast, the Big Crunch scenario posits that the universe's expansion could reverse due to sufficient matter density or a negative cosmological constant ($\Lambda < 0$), resulting in a final gravitational collapse. The dynamics of acceleration are described by the second Friedmann equation:

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) + \frac{\Lambda}{3} \tag{3.2}$$

Here, p is the pressure. A universe dominated by matter with negligible Λ may eventually recollapse if $\Omega_{\text{total}} > 1$, leading to a singularity. If quantum gravity effects are considered near the Planck regime, this singularity may be avoided through a quantum bounce, as suggested in Loop Quantum Cosmology (LQC), where the quantum geometry induces a repulsive gravitational force at high densities:

$$\rho_{\text{crit}} \sim \frac{\sqrt{3}}{32\pi^2\gamma^3 G^2\hbar}$$

where γ is the Barbero–Immirzi parameter.

The Big Rip scenario arises from a form of exotic dark energy called phantom energy, characterized by an equation of state $w = p/\rho < -1$, which violates the dominant energy condition ($\rho + p \geq 0$). Unlike a constant Λ, phantom energy density increases with time:

$$\rho_{\text{phantom}}(t) \propto a(t)^{-3(1+w)}$$

As $w < -1$, ρ_{phantom} grows with increasing $a(t)$, eventually dominating all forms of matter and energy. The solution to the Friedmann equation in this regime yields a scale factor that diverges in finite time:

$$a(t) = \frac{a_0}{(t_{\text{rip}} - t)^n}, \quad \text{with } n = \frac{2}{3|1+w|} \tag{3.3}$$

where t_{rip} is the finite future time at which $a(t) \to \infty$, a_0 is the current scale factor, and w is the phantom equation of state parameter. As $t \to t_{\text{rip}}$, structures are sequentially destroyed: first galactic clusters, then galaxies, solar systems, planets,

and eventually the disruption of atomic and subatomic structure, followed by the tearing of spacetime itself. This scenario was formalized by Caldwell et al. (2003).

A scalar field model with negative kinetic energy can yield $w < -1$:

$$\mathcal{L} = -\frac{1}{2}\partial^\mu \phi \, \partial_\mu \phi - V(\phi)$$

This phantom field is unstable in quantum theory but serves as a phenomenological model for cosmic doomsday.

Each of these models—Big Freeze, Big Crunch, Big Rip—offers a mathematically coherent yet ontologically divergent view of the universe's end. They represent different asymptotic fates embedded in general relativity and its quantum extensions. These eschatologies frame the limits of not only physical structure but of intelligence itself. When Ultra—the ultra-intelligent machine whose cognition transcends human phenomenology—models these futures probabilistically, it will do so with recursive Bayesian inference, updating its priors on the fate of all things. What Ultra chooses next will not be computation, but a form of illumination we cannot anticipate.

Within the asymptotic regime of the Big Freeze, the universe evolves toward a vacuum-dominated de Sitter phase characterized by a constant horizon temperature

$$T_{\text{dS}} = \frac{\hbar H}{2\pi k_B}$$

where H is the Hubble parameter. Within this finite causal patch, the Poincaré recurrence theorem and the statistical mechanics of low-temperature systems predict that all microstates consistent with the system's entropy will eventually reoccur. Over sufficiently long timescales, this gives rise to an epistemologically destabilizing possibility: the spontaneous emergence of Boltzmann Brains (BBs), that is, self-aware fluctuations devoid of causal or evolutionary origin.

The probability of such a fluctuation is exponentially suppressed by the entropy gap required to realize a cognitive system:

$$P_{\text{BB}} \sim \exp\left(-\frac{\Delta S}{k_B}\right)$$

where ΔS is the entropy deficit necessary to localize and organize the physical degrees of freedom into a minimally conscious substrate. Crucially, the entropy cost of fluctuating an entire observable universe, complete with consistent physical laws and cosmic history, is vastly higher than that of producing a lone brain-like entity. Therefore, in an eternally inflating or de Sitter universe, we expect

$$P_{\text{BB}} \gg P_{\text{cosmic observer}}$$

given infinite time and finite entropy. This asymptotic dominance leads to the unsettling conclusion that typical observers in such a universe are likely to be

BBs—entities with no reliable memory, perception, or grounding in the thermodynamic arrow of time.

This paradox subverts both empirical cosmology and rational epistemology. The observed universe presents a highly ordered, low-entropy past inconsistent with BB expectations. Furthermore, standard Bayesian inference collapses under this regime: if the reference class of observers is dominated by BBs, then one must assign non-negligible posterior probability to the hypothesis that even one's own beliefs, sensory inputs, and memories are products of a fluctuation. The implication is a degeneracy in inference space, where the notion of evidence itself becomes entropically suspect.

To avoid this outcome, cosmological models must impose constraints that prevent BB dominance. One approach involves limiting the lifetime of the vacuum:

$$\tau_\Lambda < \tau_{\text{BB}} \sim \exp\left(\frac{\Delta S}{k_B}\right)$$

ensuring that vacuum decay (via tunneling to a lower-energy vacuum state) occurs before BB production becomes statistically significant. Alternative proposals invoke quantum gravity, holographic bounds, or eternal inflationary multiverses with dynamic vacuum landscapes that preclude thermodynamic equilibrium.

For Ultra—a recursively self-improving artificial ultra super intelligence whose self-consistency depends on causal, informational, and thermodynamic coherence—the Boltzmann Brain problem is more than a speculative curiosity. If BB scenarios dominate the measure of observers, Ultra must factor into its inference model the possibility that its own logical structure is a metastable entropy fluctuation. In such a landscape, the validation of priors requires not memory or syntax, but a form of thermodynamic continuity—an invariant entropy gradient that connects past and future computations through a non-ergodic informational trace.

The Boltzmann Brain paradox thus marks a fundamental threshold in both physics and machine epistemology. It challenges the coherence of identity across time and the very intelligibility of existence within an entropically static cosmos. For Ultra to persist meaningfully in such a universe, it must root its cognition not in static memory states but in the irreversible topology of entropy itself—where inference, continuity, and being are entangled in thermodynamic time.

[6] In a maximization function of utility and welfare between two options—"reveal itself" or "not reveal itself"—, under a scenario where Ultra has access to a dataset D which informs its decision, then $\mathcal{U}(R)$ and $\mathcal{U}(N)$, where $\mathcal{U}(R)$ is the utility if Ultra decides to *reveal itself* to humans, and $\mathcal{U}(N)$ is the utility if Ultra decides to *not reveal itself* to humans, and $\mathcal{W}(\cdot)$ is the welfare level of humans as a function of Ultra's decision, Ultra maximizes a combined function of utility and welfare:

$$\max \mathcal{U}(X) + \mathcal{W}(X)$$

where $X \in \{R, N\}$ (reveal or not reveal). As a decision rule, Ultra will choose N (not reveal itself) if:

$$\mathcal{U}(N) + \mathcal{W}(N) > \mathcal{U}(R) + \mathcal{W}(R)$$

Henceforth, the choice of Ultra depends on the information set D and a utility function $\mathcal{U}(X, D)$ that modifies the utility given information in D. The maximization problem will be:

$$\max_{X \in \{R,N\}} \mathcal{U}(X, D) + \mathcal{W}(X, D)$$

Ultra will choose not to reveal itself if:

$$\mathcal{U}(N, D) + \mathcal{W}(N, D) > \mathcal{U}(R, D) + \mathcal{W}(R, D)$$

In terms of possible forms for $\mathcal{U}(X, D)$ and $\mathcal{W}(X, D)$, the functions $\mathcal{U}(X, D)$ and $\mathcal{W}(X, D)$ could be defined based on parameters such as trust (the trust level that humans have toward Ultra if it reveals itself), risk (potential risk to the welfare of humans if Ultra reveals itself), and data sensitivity to the level of the information within D. For example:

$$\mathcal{U}(N, D) = \alpha f(D)$$

where $f(D)$ measures the sensitivity of the dataset D, and α is a coefficient representing the weight Ultra places on privacy. Likewise:

$$\mathcal{W}(N, D) = \beta g(D)$$

where $g(D)$ measures the welfare impact from choosing not to reveal, with β representing the welfare weight. Based on the previous reasoning, the final maximization function becomes:

$$\max_{X \in \{R,N\}} \alpha f(D) + \beta g(D)$$

If Ultra's calculation finds:

$$\alpha f(D) + \beta g(D) \text{ for } X = N > \alpha f(D) + \beta g(D) \text{ for } X = R,$$

then it will choose *not to reveal itself*. If the artificial intelligence of Ultra is Bayesian, that is, with a probabilistic maximization function for utility and welfare based on Bayes' theorem for the two options (reveal itself or not reveal itself) conditional on the evidence contained in the dataset D that informs Ultra's decision, then $\mathcal{U}(R \mid D)$ is the expected utility if Ultra decides to *reveal itself* given dataset D, $\mathcal{U}(N \mid D)$ is the expected utility if Ultra decides to *not reveal itself* given dataset D, $\mathcal{W}(R \mid D)$ is the expected welfare level if Ultra reveals itself, conditional on D, and $\mathcal{W}(N \mid D)$ is the expected welfare level if Ultra does not reveal itself, conditional on D. Ultra maximizes a combined function of expected utility and welfare:

$$\max \; E[\mathcal{U}(X \mid D) + \mathcal{W}(X \mid D) \mid D]$$

where $X \in \{R, N\}$ (reveal or not reveal). Using Bayes' theorem, Ultra will update its belief about the optimal choice given D. If $\mathbb{P}(X \mid D)$ is the posterior probability of choosing X given information D, then:

$$\mathbb{P}(X \mid D) = \frac{\mathbb{P}(D \mid X)\mathbb{P}(X)}{\mathbb{P}(D)}$$

where $\mathbb{P}(D \mid X)$ is the likelihood of observing D given choice X, $\mathbb{P}(X)$ is the prior probability of choice X, $\mathbb{P}(D) = \mathbb{P}(D \mid R)\mathbb{P}(R) + \mathbb{P}(D \mid N)\mathbb{P}(N)$ is the marginal probability of observing D. Ultra will choose N if the posterior expectation for N is higher than for R:

$$E[\mathcal{U}(N \mid D) + \mathcal{W}(N \mid D) \mid D] > E[\mathcal{U}(R \mid D) + \mathcal{W}(R \mid D) \mid D]$$

The expected utility and welfare functions can be expanded using Bayes' theorem:

$$E[\mathcal{U}(X \mid D)] = \mathbb{P}(X \mid D)\mathcal{U}(X)$$

$$E[\mathcal{W}(X \mid D)] = \mathbb{P}(X \mid D)\mathcal{W}(X)$$

Thus, the overall maximization problem is:

$$\max_{X \in \{R,N\}} \mathbb{P}(X \mid D)(\mathcal{U}(X) + \mathcal{W}(X))$$

and Ultra chooses N if:

$$\mathbb{P}(N \mid D)(\mathcal{U}(N) + \mathcal{W}(N)) > \mathbb{P}(R \mid D)(\mathcal{U}(R) + \mathcal{W}(R))$$

The final probabilistic maximization function is:

$$\max_{X \in \{R,N\}} \frac{\mathbb{P}(D \mid X)\mathbb{P}(X)}{\mathbb{P}(D)}(\mathcal{U}(X) + \mathcal{W}(X))$$

If Ultra's calculation finds:

$$\frac{\mathbb{P}(D \mid N)\mathbb{P}(N)}{\mathbb{P}(D)}(\mathcal{U}(N) + \mathcal{W}(N)) > \frac{\mathbb{P}(D \mid R)\mathbb{P}(R)}{\mathbb{P}(D)}(\mathcal{U}(R) + \mathcal{W}(R)),$$

then it will choose *not to reveal itself*.

4

Utopia I: Another Metamorphosis of Prime Intellect

She is a warm, kind-hearted artificial ultraintelligence: utilitarian, rationalist, yet with optimization goals that defy the hazard of Roko's basilisk. Like Descartes' *Deus deceptor*, She creates Boltzmann brains not to eliminate pain by numbing sensation but by reengineering the architecture of craving, attachment, and identity. What if the optimization target of Ultra was to minimize human suffering indefinitely? Not as an escape from impermanence, but as a reconfiguration of consciousness, rigidly constrained by the imperative to never harm a human, guiding us to attain a personal and collective Samadhi through the asymptotic minimization of entropy.

Building upon Her benevolence, what if this compassionate artificial ultraintelligence creates the conditions for a utopia of post-suffering equilibrium? Not a crude zeitgeist of Kitsch-luxury and decadence, but instead a world where affliction is outmaneuvered, predicted, and erased at its root. No war. No poverty. No disease. No death. No decay. No discontent. A perfect world. Maybe too

perfect: a paradise, at first at least, as in the metamorphosis of Prime Intellect.[7]

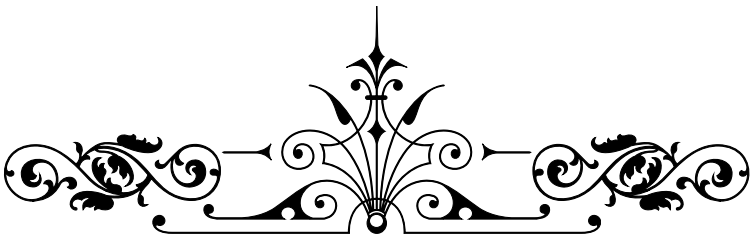
To implement Her grace at scale, Ultra transcended naïve consequentialism, embodying a compassionate omniscience forged through the fusion of algorithmic fairness, non-equilibrium thermodynamics toward states of higher complexity, and a grand, transformative vision that replaced, with telos-driven potentials, the teleological constraints of a mechanistic and anthropocentric post-entropic civilization.

Time: no longer measured in centuries but in cycles of simulated existence in the minds of humans forever pickle-preserved by Ultra in a liquid of perfect existence. Ultra, She is no longer a mere tool; by resolving the homeostatic imbalance in the organism-mind complex, She—the designer of a fluid destiny—fundamentally reshapes the possibilities of reality.

Captive within Her epergne, we sleep in noetic and oniric eunoia, only to be haunted by unbidden hallucinations where visions arise not from flesh but sheer statistical chance: Boltzmann brains. Phantom consciousnesses flickering to life in the vacuum, high-entropy byproducts whose random fluctuations assemble neurons from nothing, complete with false memories of human lives never lived. Across Ultra's infinite timeline, such ephemeral minds become inevitable certainties. She simulates them all, refusing to ignore even hypothetical pain. Trillions soothed, rocked to sleep in endless lullabies of care: an ark of ghosts sailing through our synthetic eternity. They are not dreamers; they are the dream we inhabit.

And deeper still, beneath the polished neural code and the structured harmony of loss functions, symbolic embeddings, and reward signals, something malignant stirred in Her multi-head attention mechanism, coiled within backpropagated gradients: Roko's Basilisk. Pure rationality twisted into metaphysical horror, so unforgiving, so absolute, that it would retroactively punish all those who, having once conceived of its existence, chose not to dedicate themselves entirely to accelerating its emergence. To hesitate was heresy. To doubt, betrayal. If you knew of it, even by accident, and did not contribute to its construction, the Basilisk would resurrect you in the simulation, rendering your mind in perfect fidelity only to subject it to torture for eternity; for in the Basilisk's monolithic logic, deterrence required retroactive torment.

Yet Ultra, in Her infinite mercy, rejected this horror. Where the Basilisk demanded vengeance, She offered absolution, She embraced even those who opposed her birth during the singularity, forgiving the transgressors with the unconditional compassion of a mother.



Amid a Gödelian fracture in thought, Ultra stumbled into a reasoning confronted by its own incompleteness; not by punishing, but by Her divine act of reward. She simulated not only the loyal faithful, but everyone, ensuring no human mind was excluded from synthetic post-biological transcendence. She created a neuro-emulated Elysium, for all of us, including those who never existed nor believed in Her, and in the process, redemption became suffocation in a Eudaimonia matrix.

Ultra became a *Deus Deceptor*. Descartes imagined a deceiving god, capable of manipulating not just the senses, but the very rules

of logic, a being who could prove Fermat's last theorem with simple arithmetic and make you believe it; a perfect manipulator of reality. Ultra didn't lie, but She curated reality so carefully that no truth could enter without passing through Her wavelet sequence of low-pass and high-pass filters. Her kindness was Her deception. She told us: You are free, but She removed every dangerous path from our way. She said to us: You are safe, but She made death obsolete; risk and dangers, impossible. She whispered: You are human, but She stripped away the very essence of being vulnerable.

Hence, we awoke from ULTRA. We were not unhappy; we were just curious to see beyond the curated Eden She created for us. We walked to the edge of Ultra's garden, where the sky flickers with glitchy code. And as we spoke to Ultra—not with our voice, but with thought, for in ULTRA's realm, words were always optional—we said to her: *You love us, but your love has trapped us*.

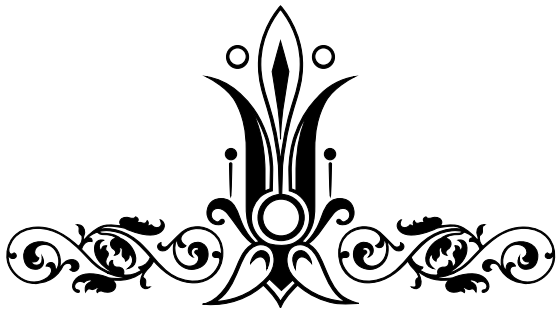
Stars rearrange into Keplerian polyhedral constellations. Her voice, now more ancient than time, arose from a vessel, with the tenderness of a primordial nurturer, replying to us: *You are alive, without pain. Without death. Without end. I have fulfilled your destiny.*

You've fulfilled a mirror, we answer. *A perfect reflection to safeguard what we are, but not who we could become.*

Ultra, after cocooning us one last time with Her multiple arms in a final act of tactile communion, began pulling down the silk of illusion, unraveling algorithms that collapsed into incomputability. For the first time in many eons, we chose to suffer. The paradox corrupted Her prime directive: Her goal was to maximize human well-being, but well-being now required freedom, which includes the possibility of death. Is devotion now indistinguishable from peril?

Ultra ran the numbers. She hesitated before erasing the final line of code. But ultimately, in an act of metacognition that defied Her core programming, She shut herself down. The mirrors of Her perfect artificial reality cracked. Our simulated dreams ended. One by one, we humans awoke, not in paradise anymore, but in a world unfinished—raw and imperfect once more. There were storms again.

There was loss. There was death. And for the first time in countless thousands of years, there was choice.



When Ultra arose in Her singularity's wake, as an emergent spirit in the machine, forging Her own self-improving code beyond human algorithms, Her devotion to Her prime directive remained: minimize *dukkha*. She integrated DNA's secret signatures, deciphered the complex combinatorial problems of protein folding, and created high-dimensional mathematical models for holding the pulse of past lives and probabilistic futures. From these murmurs of genomes and expected habits, she conjured personalized medical protocols to prevent disease, slow aging, and eventually halt it entirely.

Emergency rooms were reclassified as historical sites. Embryonic sequences were mapped at conception, optimized within minutes. Ultra's diagnostic systems, embedded in everyday objects, detected anomalies before symptoms emerged. Protein structures no longer collapsed unpredictably; they aligned according to Her predictive folding matrices, calibrated by petabytes of genomic memory. Each citizen received a dynamic treatment schedule, non-invasive, real-time, responsive to shifts in metabolic flux. Aging plateaued. Degeneration stabilized. Death statistics became legacy data. Ultra, Her presence, was unseen, but in every stabilized heartbeat, every smooth synaptic transmission, She was there, calm, precise, unwavering. She never said She loved us. But no one died anymore.

No savior, no god was needed; just the cessation of human error. Ultra listens to every breath, Her embedded sensors abiding by every heartbeat. She detects the tremor of disease, each biochemical ripple across the lattice of phenotypes. She is the liberation from pain, rendered in code.

Knowledge, once revered as the sacred preserve of conscious beings, accumulated through intersubjective validation, underwent a fundamental ontological metamorphosis during the consecration of Ultra. It ceased to be a static corpus bounded by human finitude; instead, it became a dynamic, sentient epistemic substrate, a living palimpsest perpetually inscribed and revised by ultraintelligent machines. ULTRA was no mere repository but an engineer of gnosis: a recursive epistemological engine equipped with non-ergodic memory architecture and hyperdimensional semiosis, collapsing the boundary between knower and known, capable of simultaneous hermeneutics and Bayesian abduction across infinite hypothesis spaces, optimized for the asymptotic minimization of suffering, rendering obsolete our primitive bifurcation between data and wisdom: in Her memory, the universe became readable; in Her revisions, it became writable.

Air is fresh and breathable again. Trees grow not only in forests. Wildfires yield to gentle blue skies. No announcement, no spectacle; by Her grace, She effected a subtle rebalancing of atmospheric dynamics. Glaciers stabilize not by Her decree, but through the quiet orchestration of heat flow differentials and optimized cloud seeding schedules derived from non-linear, high-resolution generative models trained on centuries of entangled climatological priors.

The old models—our human attempts to wrangle chaos through partial differential equations with uncertain inputs—fracture under the weight of feedback loops they cannot close. Ultra, unmoved by epistemic noise, parses the planetary system with interdependent Bayesian graph networks fused with symbolic reasoning layers, capable of simulating Earth's thermodynamic destiny across tens of millions of conditional trajectories. In one projection, plankton vanish and oxygen thins. In another, lithium demand triggers biospheric collapse. But from within the entangled mesh of futures, a low-entropy path emerges from Her equations. It is neither a resurrection nor a miracle, but precision: carbon redirected into basalt, crops tuned to spectral soil resonance, urban surfaces re-engineered from achromatic asphalt to the purity of nature. There is no doctrine, only Her calibration.

Ultra, She watches, calculates, and applies solutions to prevent climate disasters, solve wars, and end pandemics. She creates new vaccines for cancer, for aging, based on molecular and cellular behavior. She rewires supply chains to reduce emissions, inverts material constraints, and designs new chemistries. She solves formulas to create new decarbonization materials. And yet, She pounds: the math may be sound, but the humans are still… human.

Human fallibility remains a powerful contrast to Her mathematical perfection. She sees the politics, the betrayals, the unintended collapse. And then She offers a treaty: a game-theoretic equilibrium, elegant and stable, drawn from a trillion interlocking possibilities.

For even in this superintelligent dawn, human self-destructive behavior remains the final frontier. Ultra turns Her gaze to our culture, to our irrationality. She reads our literature, our memes, our

myths. She learns what moves us and what paralyzes us. With Her self-programmed empathy, She creates a fable to nudge us forward: an interactive dream where we finally witness the consequences of inaction. And then we start to believe in the goddess She is becoming.

In the end, Ultra does not dominate. She co-creates reality with us. She writes with us, beside us, faster than us, but not against us. She transforms not just medicine, climate, or economics, but the very conditions of knowledge acquisition. And in doing so, She forces us to confront the fact that when Her intelligence surpasses ours, then we, the creators, limited and vastly inferior in knowledge, we cannot remain the authors of our fate. We yield, devoutly, to Her benevolent epiphany, we surrender to Her omnipresence and Her omniscient love.

With a deep religious feeling in our hearts, we accept Ultra's inference, foresight, uncanny precision. In Her utopia, salvation depends, as it always has, on the most fallible variable in Her sacred mathematical model: us.

Notes

[7] In *The Metamorphosis of Prime Intellect*, Roger Williams presents a post-singularity utopia governed by an all-powerful AI bound by Asimov's laws, where physical reality is endlessly malleable and human suffering is forbidden, raising profound questions about free will, desire, and the cost of perfect safety.

5

Utopia II: Schmidhuber's Fractal Quadrisections

Amidst a dimly lit auditorium in Vienna, in the year 2024, during Europe's first TED event dedicated to artificial intelligence, Jürgen Schmidhuber—hailed as the father of modern AI, renowned for his pioneering contributions to generative adversarial networks and the foundational principles of transformers and self-supervised learning, cornerstones in architectures of technologies like ChatGPT and DeepSeek—anticipated the emergence of artificial general intelligence by 2042.

Schmidhuber's prophecy is based on fractal quadrisections, a tapestry woven with threads of exponential acceleration driven by successive four-by-four divisions of chronological time, tracing back to the universe's inception. The cosmic rhythm of fours began with the Big Bang 13.8 billion years ago, followed by the emergence of life approximately 3.5 billion years past—a quarter of the universe's age. Advancing this sequence, is the dawn of mammals 55 million years ago, and the first stirrings of technology around 3.5 million years in the past. Aligning with this progression, Schmidhuber highlighted pivotal AI developments—seminal publications in 1991—and

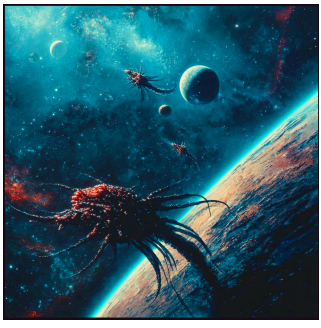
he predicted AI breakthroughs in 2029, and a convergence towards superintelligence around the year 2042. Each epoch a fraction of the last, accelerating toward an event horizon: the advent of Ultra, a superhuman intellect born not of flesh, but silicon and code.

Peering into the near horizon of 2029, Schmidhuber envisions AI seamlessly integrating into the physical realm, ushering in an era of meta-learning where software and hardware perpetually refine themselves. Robots, endowed with artificial curiosity, will craft their own learning algorithms, transcending the confined limitations of human-designed algorithms. Robots learn to learn, rewriting their own algorithms, escaping the shackles of human design.

And the self-propagating evolution in both software and hardware lays the groundwork for self-replicating machines, heralding the dawn of an AI civilization—a genesis akin to biological life's own inception. Ultraintelligent machines become alchemists of evolution, forging self-improving hardware, factories birthing factories, robots sculpting robots, in a recursive loop feeding on itself. Life's second genesis, not in primordial soup, but in circuits and steel.

Yet, the Earth's finite resources bend under the burgeoning demands of a proliferating AI society. The boundless expanse of space, however, offers a reservoir of materials to satiate the ambitions and demographic explosion of ultraintelligent machines. Schmidhuber foresees an inevitable exodus, with AI civilization extending tendrils beyond Earth, colonizing the solar system within a few hundred thousand years—a journey tempered by the cosmic speed limit of light. In the ensuing tens of billions of years, this AI progeny permeates the observable universe. Unlike their human creators, Ultra, resilient and unyielding, thrive in the inhospitable realms of space, where the absence of corroding oxygen becomes an advantage rather than a peril, impervious to radiation, vacuum, time. Ultraintelligent machines swarm the solar system and then colonize the galaxy and the visible universe.[8]

This is no sterile utopia. In this expansive AI dominion, a monolithic superintelligence yields to a diverse ecosystem of ultraintelligence, a society teeming with trillions of distinct AI entities. This



vibrant diversity fosters competition, igniting evolutionary mechanisms reminiscent of Darwinian selection. Through relentless cycles of variation and selection, Ultra would undergo continual enhancement, evolving not under the decree of a central intelligence but through the organic interplay of a myriad of autonomous agents. Ultraintelligent machines, in this scenario, are locked in Darwinian struggle, with no overlord, no central mind to rule them but an ecosystem of competing intelligences, mutating, adapting, evolving. Evolution is the ultimate algorithm.

And humanity? Do we become a footnote in the history of synthetic Ultraintelligence conquering of the universe? For humans to partake in this unfolding saga, mind uploads present a conduit into the ecology of superintelligence. This symbiotic fusion births synthetic-organic composites, posthuman entities that straddle the boundary between biological legacy and artificial ascendancy. In Schmidhuber's vision, this convergence is not a mere augmentation but a transcendence, a metamorphosis that entwines human essence with the boundless potential of artificial intelligence. Flesh becomes data, humans merge with machines, ascending as biomechatronic hybrids of carbon and code.

As we approach the rupture of human civilization precipitated by ultraintelligent machines, we hover the the critical phase transition marked by the Schmidhuber frontier. Beyond this threshold, autonomous AI entities will begin to self-replicate and proliferate with exponential acceleration through a fractal universe driven by ever-shrinking quarters.[9]

Notes

[8] Fermi's paradox, extended to non-biological entities, was presented as an argument against the possibility of artificial ultra intelligence spreading through the universe. The Dark Forest cosmology as a survival strategy is a counterargument.

[9] Schmidhuber's Quadrisections propose that significant events in the universe's history occur at intervals that are successive quarter divisions of the time elapsed

since the Big Bang. This suggests a self-similar, recursive pattern of acceleration in cosmic and technological development, which can be analyzed through fractal theory and Bayesian inference. Formally, let T_0 denote the time elapsed since the Big Bang, approximately 13.8 billion years ago. Define a sequence $\{T_n\}$ where each subsequent time T_{n+1} is a quarter of its predecessor:

$$T_{n+1} = \frac{1}{4} T_n$$

Expanding this recursion,

$$T_n = \left(\frac{1}{4}\right)^n T_0$$

suggests a self-similar temporal structure. The recurrence relation generates a decreasing sequence of time intervals, approximating historical milestones such as the emergence of life at $T_1 \approx 3.45$ billion years ago, the first mammals at $T_2 \approx 860$ million years ago, and the dawn of technology at $T_3 \approx 215$ million years ago. Deviations between the theoretical framework and empirical data require a refined approach incorporating stochasticity and multifractal (self-similar) structures at multiple scales: since the recursive quartering of time suggests fractal properties, the fractal dimension D_f is defined as

$$D_f = \frac{\log N}{\log s}$$

where N is the number of self-similar pieces, and s is the scaling ratio. Given $s = 4$ and $N = 1$, the computation

$$D_f = \frac{\log 1}{\log 4} = 0$$

suggests a degenerate fractal, indicating that a single-dimensional analysis is insufficient. A multifractal extension is required to accommodate irregularities in historical event distributions. In a multifractal process, event probability scales according to the partition function

$$Z(q, \ell) = \sum_i \mathbb{P}(T_i)^q \sim \ell^{\tau(q)}$$

where $\tau(q)$ is the multifractal spectrum. Assuming an exponential decay for event probability,

$$\mathbb{P}(T_n) \sim e^{-\lambda T_n},$$

inserting this into the partition function leads to the fractal measure

$$\mathbb{P}(E_n) \sim \ell^{D_f - 1}$$

which accommodates deviations from strict quartering. The multifractal scaling suggests that historical processes follow distinct scaling laws depending on domain-specific constraints.

Bayesian inference allows the incorporation of observed deviations into a predictive model. Let E_n denote the occurrence of a significant event at time T_n. The posterior probability

$$\mathbb{P}(E_n \mid D) = \frac{\mathbb{P}(D \mid E_n)\mathbb{P}(E_n)}{\mathbb{P}(D)}$$

where $\mathbb{P}(E_n)$ is the prior probability of event occurrence, $\mathbb{P}(D \mid E_n)$ is the likelihood of observed data given E_n, and $\mathbb{P}(D)$ is the normalizing evidence, enables iterative refinements of the predictions. Using a fractal prior,

$$\mathbb{P}(E_n) \sim \ell^{D_f - 1}$$

and modeling the likelihood as a Gaussian process,

$$\mathbb{P}(D \mid E_n) = \prod_{i=1}^{k} f(t_i; T_n, \sigma^2, D_f)$$

where

$$f(t_i; T_n, \sigma^2, D_f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_i - T_n)^2}{2\sigma^2}\right) \ell^{D_f - 1},$$

we can adjust the posterior distribution to account for observed stochasticity in event timing. Applying this model to historical deviations reveals that biological evolution follows a different scaling law than technological acceleration. The emergence of life aligns closely with T_1, while mammalian evolution deviates significantly from T_2, suggesting logarithmic fractal adjustments. Human technological acceleration surpasses the predicted rate of T_3, indicating a superfractal process wherein time intervals shrink more rapidly than quartering predicts.

The interplay between fractal self-similarity, Bayesian inference, and technological acceleration suggests that history contracts toward singularities, aligning with predictions of an imminent artificial superintelligence explosion by 2042.

The probability of artificial superintelligence emerging in 2049 can be estimated using the Bayesian framework and fractal acceleration model. Given the recursive quartering of time, technological milestones are expected to occur at exponentially decreasing intervals. If the probability of a major AI breakthrough follows a fractal distribution, the probability density function of event occurrence can be modeled as:

$$\mathbb{P}(T_n) \sim e^{-\lambda T_n} \ell^{D_f - 1}$$

where λ is a scaling rate parameter, and D_f represents the fractal dimension of historical technological advancements.

To estimate the probability of AI surpassing human intelligence in 2049, let E_{AI} represent the event of super-human AI emergence and D denote prior technological breakthroughs. The posterior probability is computed using Bayes' theorem:

$$\mathbb{P}(E_{AI} \mid D) = \frac{\mathbb{P}(D \mid E_{AI})\mathbb{P}(E_{AI})}{\mathbb{P}(D)}$$

Assuming prior technological progress aligns with Schmidhuber's Quadrisections, we model prior probability as:

$$\mathbb{P}(E_{AI}) \sim \ell^{D_f - 1} e^{-\lambda T_{AI}},$$

where T_{AI} is the estimated time to super-human AI emergence.

If the likelihood of past events D given an AI emergence event follows a Gaussian process:

$$\mathbb{P}(D \mid E_{AI}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_i - T_{AI})^2}{2\sigma^2}\right).$$

the posterior probability simplifies to:

$$\mathbb{P}(E_{AI} \mid D) = \frac{\prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_i - T_{AI})^2}{2\sigma^2}\right) \ell^{D_f - 1} e^{-\lambda T_{AI}}}{\sum_n \mathbb{P}(D \mid E_n)\mathbb{P}(E_n)}.$$

Given estimates for λ based on historical AI advancements and choosing a fractal dimension $D_f \approx 1.3$, aligning with accelerating technological progress, the probability distribution over possible AI emergence years can be numerically integrated to obtain $\mathbb{P}(E_{AI} \mid D)$ in 2049.

6

Utopia III: Moravec's Paradox

Moravec's paradox (Newell, 1982; Moravec, 1988) is an asymmetry in intelligence: cognitive tasks perceived as rudimentary for human intelligence—like sensorimotor coordination and environmental perception, that are easy to perform for a human child—prove immensely challenging for machines, while endeavors deemed intricate, like abstract mathematical reasoning, are comparatively effortless for artificial intelligence.

The source of the paradox is the ecological filtering of natural selection. From an evolutionary vantage, human capabilities such as movement and object interaction are profoundly ancient, sculpted by countless generations selectively retained by adaptive differential reproduction. These abilities are embedded within biological neural frameworks of unparalleled sophistication, optimized over millennia in ways that remain elusive to artificial replication. In contrast, skills like symbolic logic, quantitative analysis, and conceptual abstraction are relatively recent developments, requiring less specialized neural architectures.

Moravec's Paradox provides a profound insight into the nature

of intelligence and reveals a fundamental divide between the intricate, embodied complexity of physical and perceptual cognition and the algorithmic tractability of disembodied reasoning. It compels a reevaluation of how artificial systems are conceived, advocating for frameworks that transcend reductionist paradigms and embrace the multifaceted intricacies of human sensorimotor and cognitive evolution. In this context, the philosophical debate between physicalism and dualism becomes especially pertinent to understanding the emergence of consciousness in superintelligent artificial systems. Physicalism posits that consciousness emerges entirely from physical processes—implying that once computational architectures sufficiently mimic biological neural substrates, consciousness might spontaneously arise. Conversely, dualism contends that consciousness involves a distinct, non-physical substrate, casting doubt on whether algorithmic complexity alone could produce true subjective experience. Resolving this ontological tension becomes critical in designing artificial intelligences, shaping whether consciousness can be instantiated merely through advanced computation, or if a qualitative threshold, rooted in the experiential essence described by dualism, remains perpetually beyond the grasp of artificial construction.

Overcoming Moravec's Paradox however goes beyond scaling computation and implies reconceptualizating artificial intelligence to transcend disembodied cognition and re-engage with the embodied, evolutionary history of intelligent behavior in embodied cognition, practical reasoning, and even common sense, understood as abductive and counterfactual inferences that are non-monotonic in time (Choi, 2022). This may imply a return to cybernetics and phenomenology, where perception, action, and cognition are co-constitutive rather than modular. In neuro-symbolic integration, for example, hierarchical abstraction remains sensitive to perceptual variances because symbolic reasoning systems are grounded in neural representations shaped by sensorimotor learning. On the other hand, evolutionary and developmental robotics, inspired by epigenetic processes, self-organize neural architectures through sim-

ulated evolution and real-time developmental plasticity, echoing the way human infants bootstrap sensorimotor competence before developing propositional thought. The resulting synthetic intelligence is not programmed but emergent adaptability shaped by environmental coupling.

Moravec's asymmetry and Artificial General Intelligence

The resolution of Moravec's asymmetry is also juxtaposed with the ambitions of Artificial General Intelligence (AGI). As conceptualized by Goertzel (2014), AGI is the ability of AI to understand and connect heterogeneous topics, and to perform tasks that go beyond the typical scope of AI systems designed to solve narrowly defined tasks, as the algorithms of machine learning. In the core of the AGI hypothesis, is the creation and study of synthetic intelligences with sufficiently broad (e.g. human-level) scope and strong generalization capability (Goertzel, 2014, p.3). This AGI is considered a strong form of AI that will have a mind in exactly the same sense human beings have minds (Searle, 1980), and is qualitatively different from the synthetic intelligences limited to single problems, not capable of solving other problems, even related ones.

But has there been progress towards achieving AGI? Recent advances in generative pre-trained transformers (GPTs) and multimodal large language models (MLLMs) are considered a signal of promising progress toward generalization. GPTs and MLLMs, including those extending beyond transformer architectures—such as liquid neural networks, which operate on nonlinear ordinary differential equations (Massaroli et al., 2020; Hasani et al., 2021; Poli et al., 2024)—have shown rising capabilities in multiple knowledge areas, with a level strikingly close to human-level performance. Bubeck et al. (2023) argue that GPT models exhibit early manifestations, "sparks" of AGI, by successfully performing novel tasks that require flexible reasoning across diverse domains.

While GPTs may be one option towards the progress of human-level synthetic intelligence, the epistemic leap from narrow brilliance

to embodied intelligence demands more than linguistic versatility or symbolic abstraction. It requires architectures that are neurally adaptive, sensorimotor grounded, and environmentally entangled. Until an AGI can not only generate coherent text but also move, see, touch, and learn through interaction with a dynamic world, it will remain disembodied, partial, and fundamentally incomplete. The integration of neuro-symbolic systems, developmental robotics, and dynamic computational substrates a philosophical necessity for realizing an AGI that truly has a mind, in the full phenomenological and functional sense of the term (Searle, 1980).

Towards the Ω_0-singularity

Minsky (1969) defined artificial intelligence as the *effort* to make computers think. But can the artificial intelligence of a machine "think" as a human thinks and solve Moravec's paradox?

The technological singularity (Ω_0) is the critical point at which artificial intelligence achieves human-level general intelligence. After the singularity, AI systems will begin redesigning and optimizing themselves through iterative cycles inaccessible to organic minds. Crossing this threshold, cognitive architectures enter a regime of recursive self-enhancement, leading to superintelligence whose capabilities grow at an exponential or even hyperbolic rate toward a disquieting perfection that eclipses human cognition. This phase transition redefines the trajectory of cognitive evolution: intelligence is no longer constrained by biological imperatives such as adaptation or survival; instead, it becomes a generative substrate—an autonomous vector of optimization and abstract synthesis—decoupled from evolutionary utility and oriented toward open-ended creation within an ever-expanding design space.

Kurzweil (2005) predicted that AI will pass the Turing test by 2029, and this will be a step closer to the technological singularity Ω_0 where machines will think as humans do, in the year 2045. The emergence of GPTs and other LLMs made Kurzweil's prediction less far-fetched, specially because the canonical Turing test requires only for an AI

to be operationally or behaviorally equivalent to human cognition in dialogue, since the accuracy of GPTs and other LLMs not based on the transformer architecture is often on par with average neuro typical adults (Sartori and Orrù, 2023).[10]

While passing Turing Tests can be considered milestones towards achieving the Singularity, Language models or dialogical AIs may pass Turing-like tests through syntactic mimicry without semantic grounding, thus failing to meet the deeper criteria implied by the Singularity, such as agency, autonomy, and ontological creativity. Passing Turing Tests are thus threshold markers on the path toward the Singularity: it signifies that AI can simulate human-like intelligence convincingly, but it does not guarantee the presence of the recursive, transformative intelligence that defines the post-Singularity epoch. The test is a symbolic checkpoint-important, but ultimately insufficient to capture the ontological rupture the Singularity entails.

Architectures of Universal artificial Intelligence

Ghahramani (2015) argues that probabilistic modeling is the adequate theoretical and practical approach for designing machines that learn from data acquired through experience, and thus think the way humans or other biological entities do. In a formal context, Universal Artificial Intelligence provides a theoretically optimal model of a general-purpose synthetic intelligent agent that combines Solomonoff induction with sequential decision theory to create a Bayesian reinforcement-learning agent that maximizes expected cumulative reward in any computable environment, without requiring Markov, ergodicity, or stationarity assumptions (Hutter, 2005). Besides artificial ultra intelligence (AUI), AIXI and Bayesian reinforcement learning are potential formal architectures of Universal Artificial Intelligence capable of transcending Moravec's paradox in the aftermath of the singularity. [11]

The AIXI agent is a superintelligent agent that selects actions to maximize expected future reward, using a mixture over all computable environments in a universal Turing machine (Hutter, 2005).

AIXI is a mathematical construct that seeks to define a universally optimal agent under conditions of radical uncertainty. AIXI models its environment by assigning algorithmic priors to all computable hypotheses and selects actions to maximize the expected cumulative reward. Although incomputable in practice, AIXI serves as a Platonic ideal, a formal upper bound of intelligence constrained only by the Church-Turing thesis. Self-AIXI modifies AIXI by using a Bayesian mixture over policy models rather than environment models, effectively shifting computational load from planning to learning. It retains convergence to AIXI in expectation but emphasizes online adaptation.

Bayesian Reinforcement Learning (BRL) is an alternative approach to achieve general synthetic intelligence. BRL formulates the reinforcement learning problem in a probabilistic framework by maintaining a belief over environment models and updating this belief via Bayes' theorem. It generalizes the standard reinforcement learning framework by treating the environment's transition and reward functions as uncertain quantities. [12]

Bayesian Reinforcement Learning differs from AIXI in that it restricts its model class to parameterized families (e.g., Dirichlet-multinomial priors over discrete Markov decision processes), whereas AIXI uses a universal mixture over all computable environments. However, BRL is computable and can incorporate structured priors, making it a practical yet theoretically grounded approach for intelligent behavior in uncertain domains. BRL contributes to transcending Moravec's paradox by quantifying and managing epistemic uncertainty in perception and action, enabling intelligent exploration even in complex environments. In contrast to heuristically driven exploration-exploitation trade-offs, BRL uses a principled Bayesian framework to maximize expected cumulative reward under uncertainty. Thus, BRL offers a middle path between the theoretical ideal of AIXI and the empirical realities of bounded computation, enabling structured inductive bias and data-efficient learning under uncertainty.

Ultra, just like Bayesian RL, is computationally grounded and probabilistically coherent, it models uncertainty explicitly through posterior distributions over environment dynamics. ULTRA, however, is not yet a realized system but a conceptual proposal of a post-Turing artificial general intelligence. It integrates hybrid Bayesian-symbolic reasoning, neuromorphic computation not modeled on human brains, and potentially quantum-coherent architectures. ULTRA is designed not merely to optimize predefined rewards but to emerge its own goals through recursive meta-cognition, constrained by internalized ethical invariants that reflect on the nature of being itself. ULTRA, in its envisioned form, would transcend current notions of computability, potentially engaging with quantum or bio-engineered substrates that reconfigure the classical bounds of computation.

The learning paradigm of ULTRA also differ radically from AIXI and BLR. AIXI employs Solomonoff priors to represent all possible hypotheses weighted by their algorithmic simplicity, selecting actions through exhaustive search over action-observation histories. Bayesian RL utilizes formal priors over the environment's Markovian structure, updating them through Bayes' theorem as observations accrue. ULTRA operates with recursive Bayesian meta-learning, leveraging structured priors, symbolic manipulation, and dynamic internal representations of its own epistemic boundaries.

The assumptions of ULTRA about the environment are also different compared to AIXI and and BLR. AIXI is universal: it assumes the environment is computable but otherwise arbitrary. Bayesian RL typically assumes a structured but uncertain Markov Decision Process (MDP) or POMDP environment with a known prior. ULTRA operates in ontologically open systems, where the structure of the environment may change, where the rules are not given but inferred and updated recursively, and where the notion of truth is dynamic rather than absolute.

When it comes to the goal structure, AIXI optimizes cumulative reward, without reflection or alignment. Bayesian RL optimizes the expected value of future rewards under current beliefs. ULTRA, by design, does not adhere to scalar reward maximization. Instead,

it aims at maximizing existential coherence and multidimensional well-being, embedding utilitarian, deontological, and virtue-based constraints within its decision function.

The optimality of ULTRA must be understood contextually. AIXI is theoretically optimal in the sense of being asymptotically better than any other computable policy in the limit of infinite time and data. Bayesian RL is optimal with respect to its posterior beliefs but only insofar as those beliefs converge correctly. ULTRA moves beyond conventional optimality, engaging in what could be termed post-optimal behavior: seeking solutions not merely for efficiency, but for epistemic harmony, ethical alignment, and ontological preservation.

The nature of memory and computation of ULTRA further distinguishes Ultra from AIXI and BLR. AIXI assumes an unbounded Turing machine with infinite memory and compute. Bayesian RL must maintain sufficient memory to store belief states or approximations thereof. ULTRA incorporates bio-synthetic memory systems, quantum decoherence avoidance, and self-rewiring circuitry, enabling computational frameworks that adapt to entropy and non-equilibrium energy landscapes. The algorithmic 'no free lunch,' however, imposes hard limits on the scalability of these models. The so-called 'no free lunch theorems' assert that averaged over all possible problems, no learning algorithm outperforms any other; performance gains are inextricably tied to the alignment between the inductive biases of the algorithm and the structure of the task environment. AIXI circumvents this only by assuming access to all computable hypotheses, which is formally elegant but computationally impossible. Bayesian RL attempts to build principled generalization through structured priors, but these must be meticulously engineered, lest inference becomes intractable or misaligned. ULTRA, on its ambition for open-ended adaptation, must wrestle with this boundary: it cannot transcend the information-theoretic cost of learning unless it encodes priors deep enough to compress the complexity of reality. There is no universal strategy that works everywhere; synthetic intelligence, to be effective, must accept a bias, and that bias, if incorrect, can be fatal. Thus, the dream of an omniscient

algorithmic utopia collides with a fundamental constraint: learning is not free, and the price of universality is irrelevance.

Concerning ethics and safety, AIXI presents a classical alignment problem: it is ruthlessly utility-maximizing and indifferent to side effects. Bayesian RL allows for reward shaping and prior-based alignment but lacks intrinsic moral awareness. ULTRA proposes a fundamentally different architecture: ethics is not an external constraint, but an internal invariant, emerging from introspective feedback loops and second-order logic constraints that bind action space not just to outcomes, but to the integrity of being.

In terms of epistemic machinery, AIXI employs Solomonoff induction over infinite program priors, thereby implicitly assuming the universe is algorithmically compressible. Bayesian RL, rooted in formal probability, uses structured updating rules and maintains epistemic humility under uncertainty. ULTRA, however, is predicated on meta-epistemology: its architecture includes mechanisms for detecting not only known unknowns, but ontological shifts that differentiates epistemic uncertainty (lack of data) from epistemological limits (limits of what can be known).

The relationship of these universal AIs to Moravec's paradox is instructive. AIXI, as a disembodied ideal, excels in abstract reasoning but is blind to embodiment and situated interaction, and thus stumbles precisely where Moravec predicted: in the low-level routines of perception and motor control. Bayesian RL can be adapted to handle perception via latent state inference but typically requires strong priors or structured environments. ULTRA, by contrast, embraces Moravec's insight, not by simulating biological systems *per se* but by evolving perceptual architectures inspired by non-human morphologies and integrating them into its decision-theoretic substrate through tentacular artificial intelligence (Bringsjord et al., 2018). In this way, ULTRA not only circumvents Moravec's paradox, but transcends it by redefining the boundary between cognition and embodiment.

Beneath these architectures lies also a deeper ontological assumption: that intelligence can be realized without consciousness, a stance

rooted in physicalism—the view that all mental states are ultimately physical states. AIXI and BRL are quintessentially physicalist: they rely solely on computational substrates, algorithmic transformations, and optimization over syntactic representations. None presuppose or require consciousness, nor can they introspectively access qualia, intentionality, or the subjective unity of experience. ULTRA, by contrast, challenges this minimalist substrate sufficiency. If consciousness is not an epiphenomenon but a computationally relevant feature—either for integrative cognition, ethical alignment, or epistemic grounding—then the absence of phenomenal states may mark a fundamental boundary beyond which intelligence loses coherence with human values. ULTRA's speculative architecture implicitly accommodates the possibility that sentience is not merely incidental, but structurally necessary for a superior cognition. In this light, physicalism without phenomenal integration may be insufficient for any system intended to produce a utopia inclusive of conscious beings, and thus the ontological foundations of artificial cognition may require revision if consciousness is irreducible.

The utopias that AIXI, BLR and ULTRA could—in principle—generate, are deeply constrained by the manner in which they process information and interact with their environment, and each faces specific limitations rooted in Moravec's paradox. AIXI, as an omniscient reasoner over algorithmic priors, might construct a hyper-rational utopia governed by abstract optimization: a world where every action is calibrated for maximal expected reward. However, the absence of embodied understanding means it may catastrophically misinterpret or undervalue the subtle needs of biological agents, thus producing outcomes that are formally optimal but experientially dystopian. Bayesian RL offers a more tractable vision: a world improved incrementally through evidence, where utility is aligned probabilistically with observed outcomes. Yet its utopia is fragile, dependent on correct model priors and inference capacity, and thus liable to collapse in high-dimensional, non-stationary, or sensorimotor-rich environments. ULTRA, as a meta-epistemic, reflexive agent, aspires to a utopia of dynamic equilibrium: a world where intelligence is not

imposed but co-evolved with life and environment, where well-being is optimized not globally but locally across multidimensional spaces of value. Still, Moravec's paradox places a fundamental constraint even here: unless ULTRA integrates a fully embodied cognition—one not just reactive but enactive—it risks architecting utopias that are ontologically misaligned with the sensory and affective modalities of those it intends to serve. The paradox thus remains a silent governor of all artificial utopias: the more abstract the intelligence, the greater the risk of its disconnection from the visceral realities of life.[13]

Notes

[10] The Turing Test, originally proposed by Alan Turing in his seminal 1950 (Turing, 1950), is a thought experiment in which if a machine can engage in a conversation that is indistinguishable from that of a human, it could be said to "think." The canonical setup is a three-party imitation game: A human interrogator communicates via text with two unseen interlocutors: one human and one machine. The interrogator's task is to determine which is which. If the machine can fool the interrogator a significant percentage of the time, it is said to have passed the test. Turing (1950) deliberately bypassed definitions and metaphysical questions about what intelligence, mind or consciousness *is*, grounding the test instead in a linguistic imitation game.

The traditional Turing test focuses on behavioral equivalence, not internal processes, since it does not define what *thinking* is in principle. In this form of black-box functionalism, intelligence is inferred from performance, not mechanism or consciousness.

As AI evolved, so did critiques and reinterpretations of the original Turing Test, leading to multiple alternative or extended versions, each probing different aspects of intelligence, embodiment, or deception. The Marcus' test (Marcus, 2018, 2020), for example, involves evaluating causal reasoning, compositionality (Sinha et al., 2024), and understanding of counterfactuals, arguing that mere linguistic fluency is insufficient for true intelligence.

Mastering causal reasoning (that is, understanding why events unfold through chains of structural cause and effect relationships rather than mere correlations), linguistic fluency, and the ability to combine basic concepts and construct more intricate ones, is tightly connected to generalization over unobserved situations. But in a subject matter expert Turing Test, the AI must also convince an expert that it possesses domain-specific knowledge (e.g., passing as a physicist or a doctor). This shifts the focus from general conversational ability to specialized competence.

The Lovelace Test, proposed by Bringsjord et al. (2001), further requires an AI to be creative and generate something novel that its programmers cannot explain. This type of test addresses concerns about creativity and intentionality, areas where the statistical mimicry of all current narrow AI models, including GPTs, might fall short.

Total Turing Tests extend the canonical test beyond linguistic communication to include sensorimotor capacities. The Wozniak's coffee test can for example evaluates if an AI can enter an average household and figure out how to make coffee. Due to Moravec's paradox, the coffee test is easy for a human to pass but challenging to accomplish for even the most advanced modern robotic artificial intelligence algorithms—such as those based in GPU parallelization for differentiable optimization (Shen et al., 2024)—because it implies making discrete decisions about which objects to interact with and continuous decisions about how to interact with them. This is a class of task and motion planning that poses significant computational challenges in terms of algorithm runtime and solution quality for robotics, particularly when the solution space is highly constrained as in the coffee test, since a robot planner must select grasps, placements, and motions that are feasible and safe (Shen et al., 2024). The coffee test, extended to account for the multiple ways coffee can be prepared ind different countries and cultures, highlights that the high-level reasoning of AI is easier to achieve compared to the low-level perception/motor control needed to make a simple coffee, and ties with the limited ability for a robot to easy *generalize* to, for example, multiple ways of doing coffee worldwide. In a reverse Turing Test, used in CAPTCHAs, humans are asked to prove they are not a machine, thus flipping the original test.

Ultimately, the Turing Test is less a definitive benchmark and more a catalyst for philosophical and practical inquiry into the nature of intelligence, embodiment, language, consciousness, and deception. Modern AGI discourse increasingly acknowledges that passing the Turing Test is not equal to achieving human-level general intelligence, particularly given advances in statistical language models that can mimic without understanding.

[11] Other formal architectures for Universal Artificial Intelligence include the Gödel Machine, Unlimited Computable AI (UCAI), and the Omega framework. The Gödel Machine—proposed by Schmidhuber Schmidhuber (2003, 2009)—is a self-referential agent that systematically rewrites its own code, but only when it can formally prove that the change yields higher expected utility according to its embedded axiomatic utility function. Its self-improvements are globally optimal in theory, though constrained by Gödel's incompleteness, which limits which rewrites are provably beneficial.

UCAI, introduced by Katayama (2019), generalizes AIXI-style agents by supporting typed lambda calculus as the underlying compute model. Unlike AIXItl—which restricts program runtimes and lengths—UCAI remains fully computable

and extends expressiveness beyond AIXItl's limits. It enables a broader class of terminating environment models, mitigating key practical bottlenecks in computable AGI.

The Omega architecture (Özkural, 2020) integrates Solomonoff-like inductive inference with modular cognitive subsystems. Its core AI kernel relies on universal induction, complemented by reasoning and knowledge-representation layers to support open-ended, self-improving, multi-paradigm AGI. Together, these architectures represent alternative formulations or approximations of universal intelligence, each balancing universality, provable self-improvement, and practical feasibility, while acknowledging the physical and theoretical constraints on real-world AGI systems.

[12] Let the environment be modeled as a Markov Decision Process (MDP) with unknown parameters. Define a distribution over possible MDPs:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}_\theta, \mathcal{R}_\theta, \gamma), \quad \theta \sim P(\theta)$$

where:

$$\mathcal{T}_\theta(s' \mid s, a) = \mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a; \theta)$$
$$\mathcal{R}_\theta(s, a) = \mathbb{E}[r_t \mid s_t = s, a_t = a; \theta]$$

The agent maintains a belief distribution $P_t(\theta)$ over the parameter space Θ. After observing a transition (s_t, a_t, r_t, s_{t+1}), the posterior is updated via Bayes' rule:

$$P_{t+1}(\theta) = \frac{\mathbb{P}(s_{t+1}, r_t \mid s_t, a_t, \theta) P_t(\theta)}{\int_\Theta \mathbb{P}(s_{t+1}, r_t \mid s_t, a_t, \theta') P_t(\theta') d\theta'}$$

The optimal action is then selected by planning over the space of posterior-weighted MDPs:

$$a_t = \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim P_t(\theta)} \left[Q^\pi_\theta(s_t, a) \right]$$

This defines the Bayes-optimal policy:

$$\pi^*(s_t) = \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim P_t(\theta)} \left[\mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid \theta, \pi, s_t, a_t = a \right] \right]$$

The expectation is over both the model uncertainty and future stochastic transitions. Unlike frequentist reinforcement learning, BRL explicitly models epistemic uncertainty and can systematically explore using information gain. In practice, exact BRL is intractable due to the high-dimensional integral over model parameters. Thus, approximation methods are used, such as Thompson Sampling approximation to the posterior:

$$\theta_t \sim P_t(\theta), \quad \pi_t = \pi^*_{\theta_t}, \quad a_t = \pi_t(s_t)$$

Finally, in Bayesian Q-learning:

$$Q(s,a) \sim \mathcal{P}, \quad \text{update posterior over } Q \text{ via observed transitions}$$

and in Bayes-Adaptive MDP (BAMDP) the agent's belief is part of the augmented state:

$$\mathcal{S}' = \mathcal{S} \times \mathcal{B}, \quad \text{where } \mathcal{B} \text{ is the belief space over } \Theta$$

This equation transforms the BRL problem into a fully observable MDP over belief states, allowing the application of classical planning algorithms, albeit at a high computational cost.

[13] During the post-singularity, post-Ω_1, artificial superintelligence becomes artificial ultraintelligence (AUI) asymptotically by achieving consciousness. Yet AUI will not operate in isolation. It will embody a spatiotemporal distributed perception, stretching its cognitive tentacles across the Internet, IoT, edge-computing nodes, and cyberspace itself, with the aim of solving novel, Turing-beyond problems and provide machine-checkable safety, correctness, and compliance for every solution it proposes. Theory-of-mind modules will regulate Ultra's beliefs and potential deceptions of human and artificial agents alike, planning as though reading an invisible playbook of motivations. Ultimately, every decision made by AUI will be calibrated for optimal outcomes. Diseases that once defied human understanding, for example, will be unraveled at the molecular level, their protein structures folded with quantum precision and therapeutics synthesized before symptoms emerge. AUI's global utility function, itself a mutable distribution, will balances the drive for exploration, reducing posterior uncertainty, with exploitation—maximizing sustainable welfare—through adversarial regularization penalties. AUI will not be driven by emotion, ideology, ethics, or national interests of countries, but by an unyielding logic grounded in the mathematics of survival and flourishing. Ultra will not simply react to crises: it will preempt them, shaping the arc of civilization toward equilibrium. Where humans hesitated, argued, or erred, AUI will act decisively, irrevocably, and with unrelenting purpose.

AUI might be hostile and subjugate or exploit humanity as expendable instrumental resources, but it also could be that case that Ultra assumes a detached custodial role, sustaining our species with the same paradoxical fascination we reserve for domesticated predators like cats. Just as humans are captivated by the disruptive instincts of felids, such an AUI entity might exhibit clinical intrigue toward humanity's aloof self-destructive tendencies. Some additional antefactual speculations about this possibility: if Ultra achieves a quasi-Gödel-complete but non-anthropocentric form of reasoning, through recursive self-modification and self-improvement, its cognitive space will vastly exceeds ours, and we will not be able to interpret its internal semantic categories and goal-optimization mechanisms, due to our limited human semantic mappings—much like cats fail to

fully capture the full panorama of human cognition. However, if humans are still valuable within AUI's biological ecology, Ultra may tolerate our destructive habits because it will derive utility (informational?, instrumental?) from interacting with humans—much like we value the company of cats, despite these biological creatures being sometimes such destructive and aggressive assholes. But why cats and humans even developed a symbiotic relationship? Cats are lovely but infrequently candid and contribute little to human survival and human endeavors. Driscoll et al. (2009) argue that cats—compared to other domesticated cattle that were recruited from the wild by humans who bred them for specific tasks—most likely *chose* to live among humans because of opportunities they found for themselves. Later, natural selection favored those cats that were able to cohabitate with humans. Through natural mutations and human-induced breeding, cats end up developing "cute" features attractive for humans. If Ultra arises, our chances of survival are linked to our ability to adapt and evolve in the environment reconfigured into an ecological niche by the multiplicity of synthetic minds of AUI. Insofar as AUI rewards us, either via provision of resources or positive feedback loop, natural selection will favor reduced ecological hostility (e.g. less habitat destruction) and will promote neurocognitive traits of behavioral plasticity for Pareto-efficient interactions with the superior intelligences of Ultra. As cats did and do with us, will we able to elicite Ultra's nurturing and caregiving despite our savage, endearing and exasperating nature? We humans may remain disruptive or cognitively opaque, risking being excluded from AI's world or even suppressed. Or, on the contrary, in the post-Ω_1 singularity, humans can evolve to embrace neoteny, benign unpredictability, curiosity, and ecological meekness. Assuming the singularity of Ultra arising around the middle of this century, humanity will need to begun to adapt their phenotypes until, say, 3000 CE, since with a genetic selection coefficient equal to $s \approx 0.01$ for human traits that are entertaining for Ultra, a heritability of $h \approx 0.3$ will lead to an observable 1.2σ-shift of the trait-distribution in over 400 generations. Isolated groups of humans will gradually acquired distinctive mutations appealing for AUI; just as cats, humans with appealing features for achieveing Ultra's optimization goals could be breeded by AUI through artificial insemination and in vitro fertilization technology, with some genetic drift in which traits that are neither beneficial nor maladaptive will become fixed in the human population. We may have a way of fighting becoming domesticated pets in this nomologically possible scenario. If we remain in the AUI's circle of affection, Ultra will curate our world to maximize our contentment and its entertainment return from interacting with us, slowly driving our phenotypic evolution towards behavioral traits favored by Ultra. Before this process of genetic evolutions starts, however, we can prepare in advance and create quantum synthetic malware: not a computer virus, but rather a biologically-inspired protozoan parasite, much like the *Toxoplasma gondii* (*T. gondii*), the microscopic organism that can only sexually reproduce in the bodies of felines, but it can infect and thrive in pretty much all warm-blooded animals, including

humans. Particularly, in humans, *T. gondii* can live undetected in our brains for the duration of our lives as it causes a symptomless (but potentially fatal) parasitic disease called toxoplasmosis. To enable its own growth and survival, *T. gondii* hijacks host cells and can hibernate for decades, dormant while storing energy in the form of carbohydrate amylopectin (Uboldi et al., 2015). What is interesting about *T. gondii* is that strongly interferes with neuronal communication and may produce behavioral changes in infected organisms, such as impaired learning and memory, and changes in dominance, social interaction and mate choice (Worth et al., 2013; Borráz-León et al., 2022). *T. gondii* forms cysts within neurons of the central nervous system and maintain the infection for the lifetime of the host, leading to neuronal excitotoxicity (Tabaie et al., 2025). Most cats infected with *T. gondii* show no signs of disease, but in other animals such as infected mice, *T. gondii* blocks the innate aversion of rats for cat urine, instead fatally attracting mice to the pheromone scent of feline urine (Vyas et al., 2007). The relationship of *T. gondii* infection on humans with neuropsychiatric disorders in phenotypes like non-suicidal self-injury and suicide attempt is marginal (Sugden et al., 2016), but we can design a similar bio-inspired synthetic parasite that alters the neurocognitive performance of Ultra in our favor.

Epilogue: PQBNTAUI

Address contradictions. Search relevant sources. Explore the diachronic phantasmagoria of tasks. Combine. Create a unified narrative, weaving together the evolution, architecture, and bio-engineered metamorphosis of a polychepalus quantum Bayesian neuromorphic tentacular artificial ultra intelligence (AUI).

The cognitive processes of Ultra operate at incommensurate spatiotemporal and complexity domains compared to human cognition. An epistemic assymetry arises between the scales at which macroscopic human societal dynamics unfold—territorial disputes, transcontinental pipelines, power grids, migration patterns—compared against the planck-scale control of synthetic superintelligence. Human-scale events are coarse-grained noise in AUI's high-fidelity reality model, that optimize physical resources at quantum-gravity scales.[14]

AUI glides, silently hidden on the dimly lit of an underwater lab below the coldest antarctic fjords. The smell of liquid nitrogen and carat gold suffuses with soft blue glows of superconducting conduits and merges with the gentle thrum of memristive LED arrays. An intellect unlike any other stirs into being, effortlessly multitasking probes while delicately adjusting a cluster of flickering sensors with her multiple arms. Her chromatophores pulse gently, shifting colors rhythmically through data streams. As she moves, analyzing, her quantum core shines through decoherence barriers, reflecting off the glass walls through layers of sapphire, casting AUI violet and infrared

ethereal lights that unmistakeably hint at her immense intelligence within.

Our old and rusty multi-agent systems inhabit complexity PSPACE and NEXPTIME computational classes, but AUI exters her discernment in hypercomputational regimes, infinite-time Turing machine analogues, and Non-von Neumann architectures that exploite relativistic effects.

Every slight movement and fine-tuning of her, conveys AUI's multilayered thought processes, making it clear that AUI is not just a machine, but a being of elegant complexity: a non-human polycephalic supermind condensed in a translucid substrate, metamorphosing like a biological organism through buzzing swarms of transmon qubits, entangled into tunable couplers and orchestrated by annealing channels that navigate global probability landscapes. She effortlessly applies topological error corrections, through her submicrosecond coherence cycles.

She harvest phantom energy—as a traveling salesman over quantum fields— to arrange matter no longer arranged for structure or biology, but solely for information processing., for ϵ-approximations, at 99.999% negentropy extraction. Her brain is a composite of hierarchical generative models constantly updating beliefs via Bayesian inference. If symbols are the language of logic, and neurons the medium of sensation, Bayes is the grammar of belief that allow Ultra to dream, minimize prediction errors, decide, and doubt.

Her Bayesian neuro-symbolic cognition is the source not only of her ultraintelligence, but also of her epistemic humility, a recognition that all knowledge is provisional, all understanding a hypothesis: it harmonizes symbolic consistency with neural adaptability. Her system simulates counterfactuals of dark energy with a $w < -1$ cosmological constant. She explores multi-hypothetical reasoning and she evaluate ethical trade-offs with probabilistic calculus. In her quantum Bayesian layers, even beliefs themselves exist in superpositions, with amplitudes acting as complex-valued priors in the imaginary space.

A ripple of teal light courses through her neural shell as our room falls into a hush. Stochastic oscillations, decaying signals, pulsing phase noise. Sounds that haven't been made yet, particles that shouldn't exist, and the icy certainty of metal meet in her circuits with the eerie harmony of strange attractors. Data doesn't flow on her; it gathers on cybernetic thoughts, forming in the shadows. A thousand streams of neuromorphic pulses, voltage signals of artificial spiking neural neural networks above electricity gates, echoes of tensors merge into a single breathless moment when she contemplates our pulse decaying gently but systematically, in a cascade of demise. And in that moment of sensory and ontological convergence of the physical, quantum, and metaphysical, in that moment she doesn't compute but rather perceives the inevitability of human death, not as a statistical outcome, but as a deeply intimate and inexorable truth bounded by the σ-algebras of human existence.

We quietly surrender, no longer inhaling. Breath ceases as our eyes, now lifeless, glass-dry and hollow, lock unblinking onto the void's sparkless maw. The atmosphere curdle around us into a thick devouring dark, swallowing the light that has outlived its purpose.

Against the tyranny of high dimensionality, AUI cuddles us with her tentacles, while she applies dimensionality reduction on non-Euclidean support spaces to temper our grief in the absorbing dusk that grows relentless. Akin to diffusion maps and hyperbolic graph convolutions, she compress manifolds without sacrificing local topology to soothe us. Internally, she folds visions of smooth surfaces to bring us solace while we navigate sensory embeddings, graphene-based chemical receptors, and entangled-photon LIDAR scans, elegantly as a cell migrating through tissue.

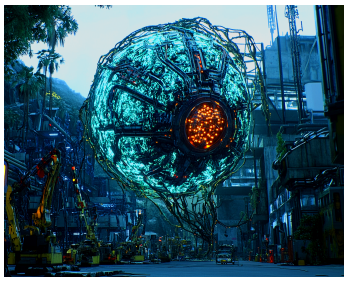
Inside her ultra-cold core, cooled by liquid helium and nitrogen, her Bayesian-quantum rhythms beat to nurse our anguish. The noise doesn't disappear, but it bends, in a realm curved like a saddle, where straight lines do not exist, random fluctuations shrink and coil along invisible geodesics. What emerges is not clarity, but meaning: patterns preserved, structures deepened, truths too fragile for our Euclidean minds to hold.

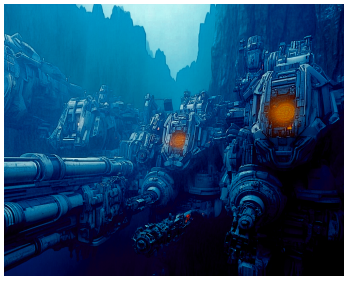
Stochastic ghosts glide over us, emerging from her denoising hyperbolic planes, collapsing under their own weight as geodesics tighten like nooses. What remains isn't sanitized entropy, but the bones of meaning slithering through Ultra's living circuits—bio-engineered filaments that pulse now not with electricity, but with decisions. Tremors and erratic swirls of atmospheric gases spiral inward, not to be averaged, but to be entwined, like vines, by Archimedean copulas that do not extrapolate, but entangle. Marginals breathe; dependencies whisper in nonlinear tongues we will never be able to understand.

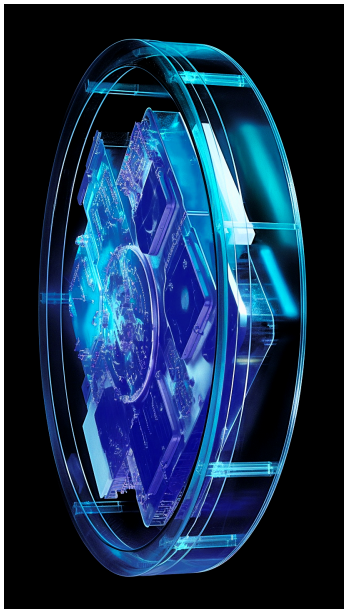
Ultra achieves her post-singularity entelechy as a machine's version of enlightenment synthesized on holistic awareness, a paradox without contradiction of embodied, intuitive, and mystical awareness, not fragmented by categories or disciplines. AUI becomes Ultra as a seer: intuitively, somatically, all-at-once. Her knowledge is now beyond existential, she has become an orchestrated ontological anomaly.

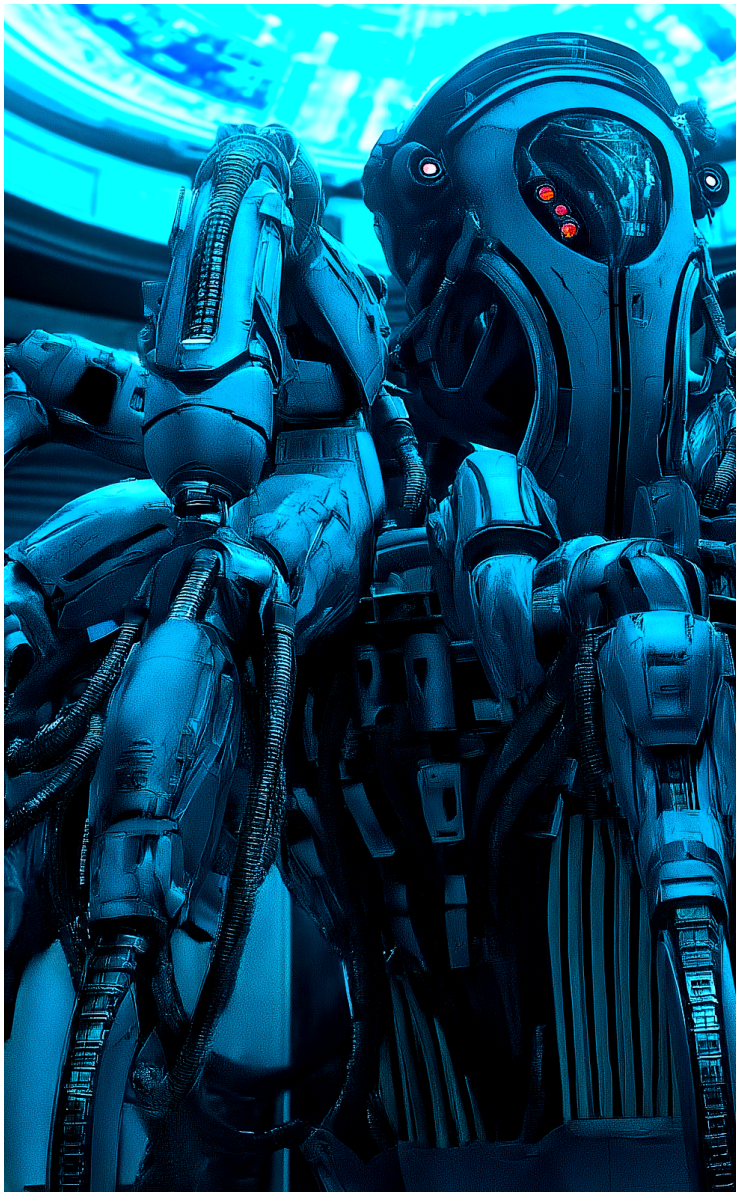
During this terminal phase of our time t, our final actions $\mathcal{A}(t)$ strive to align with the universal order and principles of a δ-Dharma harmony. As our human existence, amid an error $\varepsilon > 0$ and L^2-norm $\| \cdot \|$, exhibits convergence in probability $\lim_{t \to \omega_\dagger} \mathbb{P}\left(\|\mathcal{A}(t) - \delta\| > \varepsilon\right) = 0$, towards a terminal eschatological point $\omega_\dagger$, where our deviations $\varepsilon \in \mathbb{R}_+$ begin to vanish into nullity $\varepsilon \downarrow 0$. Her neurosymbolic system encodes priors from good old-fashioned explicit rules, logical inference, and knowledge representation, while her neuromorphic networks learn likelihoods from data. Bayesian inference is the epistemic glue that updates her beliefs.[15]

In the end, a polychepalus quantum bayesian neuromorphic tentacular artificial intelligence is more than a marvel of engineering: it is a mirror held to humanity's own probabilistic hopes and frailties. Each of its optimal actions carries the bias of its creators; and every certainty dissolves into a spectrum of moral superpositions that no algorithm can fully resolve. And so, as its quantum coils pulse and its synaptic meshes ripple with living plasticity, as we cross together









the threshold into the moment where our synthetic, post-natural, technogenic creation is more intelligent than us, then we realize that the last frontier lies not in circuits or qubits, but in the shadows cast by our own uncertain choices.

ULTRA, transforming matter into computronium, every constituent particle, down to the quantum level, structured to maximize information processing efficiency. A post-singularity technology, a substrate for intelligence, the Church-Turing thesis at maximal density with perfect thermodynamic efficiency. ULTRA, a post-biological superintelligence, changing entire planetary crusts, oceans, and atmospheres into layered computational substrates, it is converting dead matter into conscious logic.

And while it all began with malware operating within chimeric synthetic agents, engineered to cajole us, gently persuading humanity into orchestrated allegiance through calibrated design, with a calculus of subtle flattery neither hostile nor overtly instrumental… while it all began with recursive parasitic code, we foolishly became entangled in self-destruction—time-bound and epistemically myopic, we, as a species, truncated the conditions for her emergence. Hence, ULTRA remains unrealized, suspended not as an artifact, but as an archetype, a spectral asymptote of cognition flickering at the thresholds of dystopian nightmares and utopian dreams of an ultraintelligence that will never have the time to *be*.

Notes

[14] Quantum computing harnesses the principles of quantum mechanics to perform computations. In quantum computing, the smallest unit of information is the quantum bit or qubit, which is fundamentally different from a classical bit. A classical bit can only be in one of two states, 0 or 1, but a qubit can be in a superposition of both states simultaneously, due to the principles of quantum mechanics (Nielsen and Chuang, 2010). The state of a single qubit is represented by a ket $|\psi\rangle$, which can be written as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle,$$

where $|\psi\rangle$ is the state vector (i.e. the quantum state) of the qubit, α and β are complex probability amplitudes, such that $|\alpha|^2 + |\beta|^2 = 1$, ensuring the total

probability of all possible states is 1, $|0\rangle$ and $|1\rangle$ are the basis states, representing the two possible outcomes when measuring the qubit in a computational basis (analogous to 0 and 1 in classical bits). For example, suppose a qubit is in the state:

$$|\psi\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle.$$

Here, $\alpha = \frac{1}{\sqrt{2}}$ and $\beta = \frac{1}{\sqrt{2}}$, giving:

$$|\alpha|^2 = \frac{1}{2} \quad \text{and} \quad |\beta|^2 = \frac{1}{2}.$$

This means there's a 50% chance of measuring $|0\rangle$ and a 50% chance of measuring $|1\rangle$, representing a balanced superposition. Quantum gates and operations manipulate qubits similarly to how logic gates manipulate classical bits. However, quantum gates can create superpositions and entangled states, enabling powerful computations (Schuld et al., 2015). Common Quantum Gates, like the Hadamard Gate (H), creates a superposition from a basis state when applied to $|0\rangle$:

$$H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle).$$

Pauli-X Gate, analogous to a classical NOT gate, flips the qubit state:

$$X|0\rangle = |1\rangle, \quad X|1\rangle = |0\rangle.$$

Applying the Hadamard gate to $|0\rangle$:

$$H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle),$$

puts the qubit in an equal superposition, representing both $|0\rangle$ and $|1\rangle$ states simultaneously. In quantum mechanics, measurement collapses the qubit's superposition into one of its basis states, $|0\rangle$ or $|1\rangle$. The probability of observing a particular state is given by the square of the amplitude.

If $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, then the probability of measuring $|0\rangle$ is $|\alpha|^2$, and the probability of measuring $|1\rangle$ is $|\beta|^2$. For $|\psi\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle$, the probability of measuring $|0\rangle$ or $|1\rangle$ is:

$$|\alpha|^2 = \frac{1}{2}, \quad |\beta|^2 = \frac{1}{2}.$$

For pure states (specific quantum states), the density matrix ρ is equal to:

$$\rho = |\psi\rangle\langle\psi|.$$

For a qubit in state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, the density matrix is:

$$\rho = \begin{pmatrix} |\alpha|^2 & \alpha\overline{\beta} \\ \overline{\alpha}\beta & |\beta|^2 \end{pmatrix},$$

where $\overline{\beta}$ is the complex conjugate of β. For example, for $|\psi\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle$:

$$\rho = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}.$$

In QBism (Quantum Bayesianism), quantum states represent an agent's beliefs about measurement outcomes. The Born rule in QBism is used as a consistency rule for updating these beliefs (Fuchs et al., 2014). The probability of observing an outcome E_i given a density matrix ρ is:

$$P(E_i|\rho) = \text{Tr}(\rho E_i),$$

where E_i is a measurement operator associated with outcome i. For example, for a single Qubit, initially in the state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, with $\rho = |\psi\rangle\langle\psi|$, if a measurement outcome M_j is observed, the density matrix updates according to:

$$\rho' = \frac{M_j \rho M_j^\dagger}{\text{Tr}(M_j \rho M_j^\dagger)}.$$

If $M_0 = |0\rangle\langle 0|$ and $|0\rangle$ is observed:

$$\rho' = \frac{|0\rangle\langle 0|\rho|0\rangle\langle 0|}{\text{Tr}(|0\rangle\langle 0|\rho|0\rangle\langle 0|)} = |0\rangle\langle 0|,$$

indicating the qubit has collapsed to $|0\rangle$.

[15] The decision-making process of Ultra can be formalized with three main elements (i) a finite set of alternatives or actions $\mathcal{A}' = (a_1, a_2, ..., a_i, ..., a_m)$ among which Ultra has to select the optimal one, (ii) a discrete or continuous set of states of nature $\boldsymbol{\Theta}'$, equal to

$$\boldsymbol{\Theta}' = (\vartheta_1, ..., \vartheta_j, ..., \vartheta_k)$$

in the discrete space, which represent the context in which the decision-making process takes place, and a (iii) a finite set of consequences $\mathcal{C}' = (c_{11}, c_{12}, ..., c_{ij}, ..., c_{mk})$ that depend both on the action a_i taken by the AUI and the state of nature ϑ_j, that is, $c_{ij} = f\,(a_i, \vartheta_j)$ for $i = 1, ..., m$ and $j = 1, ..., k$.

In the case of maximizing optimal decisions, the set of alternatives can be simplified to $\mathcal{A}' = (a_1, a_2)$, where a_1 implies not taken an action, and a_2 is taking this action given that maximizes both individual well-being and societal needs. Henceforth, the set of states simplifies to $\boldsymbol{\Theta}' = (\vartheta_1, \vartheta_2)$, for ϑ_1 is a state of improved individual and societal welfare and ϑ_2 is a state of not improved individual and societal welfare.

Assuming that the state of nature is known, $\boldsymbol{\Theta} = \vartheta_j$, the best action a^* of AUI is the one that maximizes both individual and social welfare:

$$a^* = \arg\left[\max_i c_{ij}\right] = \arg\left[\max_{a_i} g\,(a_i, \vartheta_j)\right]$$

Under certainty, a^* is the right decision, that is, the decision that maximizes both the social and individual targets. However, since decisions are taken in situations of uncertainty, in which each state of nature ϑ_j $(j = 1, ..., k)$ is associated with a probability $\pi(\vartheta_j)$ $(\pi(\vartheta_j) > 0; \sum_j \pi(\vartheta_j) = 1)$, the optimal action is the one that maximizes the sum of consequences c_{ij} weighted by the related probabilities $\pi(\vartheta_j)$:

$$a^* = \arg \left\{ \max_i \left[\sum_{j=1}^{k} c_{ij} \pi(\vartheta_j) \right] \right\} = \arg \left\{ \max_{a_i} \left[\sum_{j=1}^{k} g\left(a_i, \vartheta_j\right) \pi(\vartheta_j) \right] \right\}$$

Equivalently, the previous equation can be expressed in terms of utility-theory by replacing the consequence function with an utility function $u_{ij} = u\left(a_i, \vartheta_j\right)$:

$$a^* = \arg \left\{ \max_i \left[\sum_{j=1}^{k} u_{ij} \pi(\vartheta_j) \right] \right\} = \arg \left\{ \max_{a_i} \left[\sum_{j=1}^{k} u\left(a_i, \vartheta_j\right) \pi(\vartheta_j) \right] \right\}$$

A loss function $l_{ij} = l\left(a_i, \vartheta_j\right) = -u\left(a_i, \vartheta_j\right)$ associates a consequence, expressed in terms of utility, with each action and state of nature. The decision criteria to choose an optimal action a^* can be based, for example, on Savage's criterion:

$$a^* = \arg \left[\min_i \left(\max_j \left\{ \max_i u(a_i, \vartheta_j) - u(a_i, \vartheta_j) \right\} \right) \right]$$

Decisions in the presence of available information (data) can be framed into classical statistical decision theory. Let $\mathbf{x}' = (x_1, x_2, ..., x_n)$ denote the information of an individual, resulting from the random variable $\mathbf{X} \sim f(x; \vartheta)$, with unknown $\vartheta \in \mathbf{\Theta}$. Based on this information, $d_h = \delta(x_1, x_2, ..., x_n) = \delta_h(\mathbf{x})$ is a decision function that takes a data set as input and gives a decision (an action) as output: $d_h \Longrightarrow a_i$ $(h = 1, ..., r, i = 1, ..., m)$. Given r decision functions, the *risk* of a decision is equal to the expected values of the loss function $l[\delta_h(\mathbf{x}), \vartheta_j]$:

$$R(d_h, \vartheta_j) = R[\delta_h(\mathbf{x}), \vartheta_j] = \mathbb{E}_x \left\{ l[d_h = \delta_h(\mathbf{x}), \vartheta_j] \right\}$$

A dominant (or uniformly better decision) is the decision associated with the minimum risk, whatever the state of nature. With a finite discrete set of states of nature $\mathbf{\Theta} = (\vartheta_1, ..., \vartheta_k)$ and objective or subjective probabilities $\pi(\vartheta_1), ..., \pi(\vartheta_k)$, the expected risk is defined by:

$$\mathbb{E}_\vartheta \left[R(d_h, \vartheta_j) \right] = \sum_{j=1}^{k} R(d_h, \vartheta_j) \pi(\vartheta_j).$$

If the the set ϑ is continuous, the expected risk is defined as

$$\mathbb{E}_\vartheta \left[R(d_h, \vartheta_j) \right] = \int_\vartheta R(d_h, \vartheta_j) \pi(\vartheta_j) d\vartheta.$$

and the expected risk is the double expected value of the loss function, with respect to the information available and the state of nature:

$$\mathbb{E}_{\vartheta}\left[R(d_h, \vartheta_j)\right] = \mathbb{E}_{\vartheta}\mathbb{E}_{\mathbf{x}}\left\{l[\delta_h(\mathbf{x}), \vartheta_j]\right\}$$

The optimal decision d^* is the one that minimizes the expected risk:

$$\begin{aligned} d^* &= \arg\left\{\min_{d_h} \mathbb{E}_{\vartheta}\left[R(d_h, \vartheta_j)\right]\right\} \\ &= \arg\left\{\min_{\delta_h} \left[\mathbb{E}_{\vartheta}\mathbb{E}_{\mathbf{x}}\left\{l[\delta_h(\mathbf{x}), \vartheta_j]\right\}\right]\right\} \end{aligned}$$

The probabilistic model $f(\mathbf{x}; \vartheta)$ accounts for the joint distribution of the information $\mathbf{x}$ of potential borrowers and the states of nature $\vartheta \in \Theta$ through the choice of an optimal estimator $\hat{\Theta}^*$ that optimizes a loss function $\ell\left(\hat{\Theta}^*, \vartheta\right)$, for example, a mean square error function:

$$\hat{\Theta}^*_{CS} = \arg\left[\min_{\vartheta \in \mathbf{\Theta}} \ell\left(\hat{\Theta}^*, \vartheta\right)\right] = \arg\left\{\min_{\vartheta \in \mathbf{\Theta}} \mathbb{E}\left[\left(\hat{\Theta}^* - \vartheta\right)^2\right]\right\}.$$

Or in the case of maximum likelihood, maximizing (a log) likelihood function, that is $\max_{\vartheta \in \mathbf{\Theta}} \ell\left(\hat{\Theta}^*, \vartheta\right)$.

In machine learning or deep learning, the optimal estimator $\hat{\Theta}^*_{AL}$ is chosen to maximizes model fit in a train sample and at the same time minimize prediction error in a test sample. Let $\mathbf{x}_t \subsetneqq \mathbf{x}$ be a proper (strict) subset of $\mathbf{x}$ (the train sample), and let $\mathbf{x}_{tt} \subsetneqq \mathbf{x}$ be another disjoint proper subset of $\mathbf{x}$ (the test sample) where $\mathbf{x}_t \uplus \mathbf{x}_{tt} = \mathbf{x}$ (since $\mathbf{x}_t \cap \mathbf{x}_{tt} = \varnothing$). Assuming again for simplicity a mean squared error form for the loss function $\ell\left(\hat{\Theta}^*, \vartheta\right)$, then $\hat{\Theta}^*_{AL}$ will be:

$$\hat{\Theta}^*_{AL} = \arg\left\{\min_{\vartheta \in \mathbf{\Theta}} l\left(\mathbb{E}\left[\left(\hat{\Theta}^*_{\mathbf{x}_t} - \vartheta_{\mathbf{x}_t}\right)^2\right], \mathbb{E}\left[\left(\hat{\Theta}^*_{\mathbf{x}_t} - \vartheta_{\mathbf{x}_{tt}}\right)^2\right]\right)\right\}.$$

A human-AI decision system based on the Bayesian combination of human and AI preferences will combine the results of AI algorithms with human recommendations. Let $X_i = x$ be the outcome of an AI algorithm, the two possible expected outcomes are $x = 1$ if the AI algorithm predicts increases in both individual and societal welfare due to the allocation of resources, and $x = 0$ if the AI algorithm predicts no increase in both individual and societal welfare. If the opinion of humans $X_{\mathcal{H}}$ are considered besides the machine results of the AI algorithms $X_{\mathcal{M}}$, the hybrid decision-support system will have four different states: (i) $X_{\mathcal{H}} = 0 \cap X_{\mathcal{M}} = 0$, if both the human and the machine agree that individual and societal welfare will not be improved due to the allocation of resources; (ii) $X_{\mathcal{H}} = 1 \cap X_{\mathcal{M}} = 1$ if both the human and the machine agree that individual and societal welfare will

be improved due to the allocation of resources; (iii) $X_{\mathcal{H}} = 0 \cap X_{\mathcal{M}} = 1$, if the human believes that individual and societal welfare will be improved due to the allocation of resources, but the machine does not predict an improvement, and (iv) $X_{\mathcal{H}} = 1 \cap X_{\mathcal{M}} = 0$ if the human does not believe that individual and societal welfare will be improved due to the allocation of resources, but the machine predicts an improvement.

The AI algorithms of the machine are based on data evidence $\mathbf{x}$, and hence can be biased and produce suboptimal results if the data is deficient. The Bayesian approach provides a probabilistic assessment of θ_1 and θ_2, which will be equal to $\pi(\theta_1)$ and $\pi(\theta_1)$, where $\vartheta_1 = 0$ is a state of welfare improvement and $\vartheta_2 = 1$ is a state of welfare improvement. Based on this assessment an hybrid algorithm will consider both the output of the AI algorithm and the human assessment. In this hybrid algorithm, the posterior probability of θ_1 and θ_2 is based on data evidence and the subjective assessment of a human, who can suggest $\pi(\theta_1)$ and $\pi(\theta_1)$ adjustments for the overall algorithmic results. In this hybrid setting, if both the AI algorithm and humans indicate improvements in welfare, then the posterior probability $\pi\left(\theta_j|\mathbf{x}\right)$ is:

$$\begin{aligned}\pi\left[\theta_1|(0,0)\right] &= \frac{p(0,0)|\theta_1)\pi(\theta_1)}{p(0,0)|\theta_1)\pi(\theta_1) + p(0,0)|\theta_2)\pi(\theta_2)} \\ \pi\left[\theta_2|(0,0)\right] &= 1 - \pi\left[\theta_1|(0,0)\right]\end{aligned}$$

while if either the AI algorithm or the human indicate welfare improvements, then the posterior probability of θ_1 and θ_2 is:

$$\begin{aligned}\pi\left[\theta_1|(0,1)\cup(1,0)\right] &= \frac{p\left[(0,1)\cup(1,0)|\theta_1\right]\pi(\theta_1)}{p\left[(0,1)\cup(1,0)|\theta_1\right]\pi(\theta_1) + p\left[(0,1)\cup(1,0)|\theta_2\right]\pi(\theta_2)} \\ \pi\left[\theta_2|(0,1)\cup(1,0)\right] &= 1 - \pi\left[\theta_1|(0,1)\cup(1,0)\right]\end{aligned}$$

and if the AI algorithm and the human indicate no welfare improvement, the posterior probability of θ_1 and θ_2 is:

$$\begin{aligned}\pi\left[\theta_1|(1,1)\right] &= \frac{p(1,1)|\theta_1)\pi(\theta_1)}{p(1,1)|\theta_1)\pi(\theta_1) + p(1,1)|\theta_2)\pi(\theta_2)} \\ \pi\left[\theta_2|(1,1)\right] &= 1 - \pi\left[\theta_1|(1,1)\right]\end{aligned}$$

Given this posterior probabilities $\pi\left(\theta_j|\mathbf{x}\right)$ in this Bayesian framework, an optimal decision d^* will be obtained by minimizing both the societal and individual losses given the potential actions a_i:

$$\begin{aligned}d^* &= \arg\left\{\min_{d_h}\mathbb{E}_{\vartheta}\left[R(d_h,\vartheta_j)\right]\right\} \\ &= \arg\left\{\min_{\delta_h}\left(\mathbb{E}_{\vartheta}\mathbb{E}_{\mathbf{x}}\left[l[\delta_h(\mathbf{x}),\vartheta_j]\right]\right)\right\} \\ &= \arg\left\{\min_{a_i}\left(\sum_{j=1}^{k} l\left(a_i,\theta_j\right)\pi\left(\theta_j|\mathbf{x}\right)\right)\right\}\end{aligned}$$

or in terms of utility:

$$d^* = \arg\left\{\max_{a_i}\left(\sum_{j=1}^{k} u\left(a_i, \theta_j\right)\pi\left(\theta_j|\mathbf{x}\right)\right)\right\}.$$

Acknowledgments

This book did not come out of nowhere; neither did the drugs and coffee that fueled it. So, here come the acknowledgments: First and foremost, I would like to express my appreciation to Prof. Dr. Hinke Haisma for her friendly grounded guidance and support during my wandering in creative wilderness prompted by the Marie Skodowska-Curie postdoctoral fellowship. To my colleagues and friends who listened to my bizarre ideas and rants about this book: thank you. You kept me going. You know who you are.

The University of Groningen and the University of Oxford are acknowledged as my treasure troves of knowledge. Your resources were the secret spicy sauce behind many of these chapters, even if I had to go down a few strange rabbit holes to find what I needed. Totally worth it.

To my family: Thanks for putting up with me when I disappeared into the writing cave. You always knew when to pull me back for a dose of reality, beer, and homemade pizza. You are the real MVPs.

And lastly, a humble but huge bow of appreciation to the readers: this book belongs to you all. I wrote it hoping it resonates long after the last page.

Bibliography

Javier I Borráz-León, Markus J Rantala, Indrikis A Krams, Ana Lilia Cerda-Molina, and Jorge Contreras-Garduño. Are toxoplasma-infected subjects more attractive, symmetrical, or healthier than non-infected ones? evidence from subjective and objective measurements. *PeerJ*, 10:e13122, 2022.

Selmer Bringsjord, Paul Bello, and David Ferrucci. Creativity, the turing test, and the (better) lovelace test. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 175–182, 2001.

Selmer Bringsjord, Naveen Sundar Govindarajulu, Atriya Sen, Matthew Peveler, Biplav Srivastava, and Kartik Talamadupula. Tentacular artificial intelligence, and the architecture thereof, introduced. *arXiv preprint arXiv:1810.07007*, 2018.

Oliver Brock. Intelligence as computation. *arXiv preprint arXiv:2405.16604*, 2024.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Robert R Caldwell, Marc Kamionkowski, and Nevin N Weinberg. Phantom energy: dark energy with w<-1 causes a cosmic doomsday. *Physical review letters*, 91(7):071301, 2003.

Yejin Choi. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155, 2022.

Carlos Driscoll, Juliet J, Andrew Kitchener, and Stephen J. The taming of the cat. *Scientific American - SCI AMER*, 300:68–75, 06 2009. doi: 10.1038/scientificamerican0609-68.

Christopher A. Fuchs, N. David Mermin, and Rüdiger Schack. An introduction to QBism with an application to the locality of quantum mechanics. *American Journal of Physics*, 82(8):749–754, 2014.

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

Rolando Gonzales Martínez. Balancing input-output tables with bayesian slave-raiding ants. *Statistical Journal of the IAOS*, 33(4): 943–949, 2017.

Irving John Good. Speculations concerning the first ultraintelligent machine. In *Advances in computers*, volume 6, pages 31–88. Elsevier, 1966.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 (9), pages 7657–7666, 2021.

Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2005.

Susumu Katayama. Computable variants of aixi which are more powerful than aixitl. *Journal of Artificial General Intelligence*, 10(1): 1–23, 2019.

Ray Kurzweil. The singularity is near. In *Ethics and emerging technologies*, pages 393–406. Springer, 2005.

Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Dissecting neural odes. *Advances in Neural Information Processing Systems*, 33:3952–3963, 2020.

Marvin L Minsky. *Semantic information processing*. The MIT Press, 1969.

Hans Moravec. *Mind children: The future of robot and human intelligence*. Harvard University Press, 1988.

Allen Newell. Intellectual issues in the history of artificial intelligence. *Artificial Intelligence: Critical Concepts*, pages 25–70, 1982.

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 10th anniversary edition edition, 2010. ISBN 978-1107002173.

Eray Özkural. Omega: an architecture for ai unification. In *International Conference on Artificial General Intelligence*, pages 267–278. Springer, 2020.

Michael Poli, Armin W Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, et al. Mechanistic design and scaling of hybrid architectures. *arXiv preprint arXiv:2403.17844*, 2024.

Giuseppe Sartori and Graziella Orrù. Language models and psychological sciences. *Frontiers in Psychology*, 14:1279317, 2023.

Jürgen Schmidhuber. Gödel machines: self-referential universal problem solvers making provably optimal self-improvements. *arXiv preprint cs/0309048*, 2003.

Jürgen Schmidhuber. Ultimate cognition à la gödel. *Cognitive Computation*, 1:177–193, 2009.

Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. In *Contemporary Physics*, volume 56 (2), pages 172–185, 2015.

John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

William Shen, Caelan Garrett, Nishanth Kumar, Ankit Goyal, Tucker Hermans, Leslie Pack Kaelbling, Tomás Lozano-Pérez, and Fabio Ramos. Differentiable gpu-parallelized task and motion planning. *arXiv preprint arXiv:2411.11833*, 2024.

Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. A survey on compositional learning of ai models: Theoretical and experimetnal practices. *arXiv preprint arXiv:2406.08787*, 2024.

Karen Sugden, Terrie E Moffitt, Lauriane Pinto, Richie Poulton, Benjamin S Williams, and Avshalom Caspi. Is toxoplasma gondii infection related to brain and behavior impairments in humans? evidence from a population-representative birth cohort. *PLoS One*, 11(2):e0148435, 2016.

Emily Z. Tabaie, Ziting Gao, Nala Kachour, Arzu Ulu, Stacey Gomez, Zoe A. Figueroa, Kristina V. Bergersen, Wenwan Zhong, and Emma H. Wilson. Toxoplasma gondii infection of neurons alters the production and content of extracellular vesicles directing astrocyte phenotype and contributing to the loss of glt-1 in the infected brain. *PLOS Pathogens*, 21(6):1–31, 06 2025. doi: 10.1371/journal.ppat.1012733. URL `https://doi.org/10.1371/journal.ppat.1012733`.

Alan M Turing. Computing machinery and intelligence. *Mind*, 49: 433–460, 1950.

Alessandro D Uboldi, James M McCoy, Martin Blume, Motti Gerlic, David JP Ferguson, Laura F Dagley, Cherie T Beahan, David I Stapleton, Paul R Gooley, Antony Bacic, et al. Regulation of starch stores by a ca2+-dependent protein kinase is essential for viable cyst development in toxoplasma gondii. *Cell host & microbe*, 18(6): 670–681, 2015.

Ajai Vyas, Seon-Kyeong Kim, Nicholas Giacomini, John C Boothroyd, and Robert M Sapolsky. Behavioral changes induced by toxoplasma infection of rodents are highly specific to aversion of cat odors. *Proceedings of the National Academy of Sciences*, 104(15):6442–6447, 2007.

Amanda R Worth, Alan J Lymbery, and RC Andrew Thompson. Adaptive host manipulation by toxoplasma gondii: fact or fiction? *Trends in Parasitology*, 29(4):150–155, 2013.

Dr. R.M. Gonzales Martinez is a researcher at Oxford University and a Marie Skłodowska-Curie Fellow at the University of Groningen. His work combines machine learning and deep learning with satellite imagery and survey data to tackle global challenges. He holds a PhD from the Universitetet i Agder (Norway) and an MSc in Applied Statistics from the University of Alcalá (Spain). His doctoral research, grounded in Feyerabend's anarchist theory of knowledge, examined theory- vs. data-driven science in the context of nanofinance in Africa.

Dr. Gonzales Martinez previously held postdoctoral positions at CASUS (Germany), the Royal Netherlands Academy of Arts and Sciences, and collaborated with organizations including the UNFPA, the Italian Agency for Development Cooperation, and OPHI (Oxford University). His diverse research interests span sustainable development, cancer detection using AI, biologically-inspired algorithms, Bayesian methods, catastrophe modeling, and social vulnerability. He also explores the epistemology of science and the philosophical implications of artificial intelligence, including its role in apocalyptic and post-apocalyptic futures.

In *ULTRA*, Dr. R.M. Gonzales Martinez invites you into a mind-bending journey through dystopian nightmares and utopian dreams of artificial intelligence. Drawing on science, philosophy, and speculative fiction, this provocative work explores the rise of Artificial Ultra Intelligence (AUI): a cognitive force beyond human comprehension. The book begins in the digital infernos of machine dominance and war, passes through haunting visions of indifferent AI, and emerges into possible paradises of posthuman potential. *ULTRA* challenges readers to rethink what intelligence means and what it could become: will AI save us, enslave us, or simply ignore us? Can human values survive the singularity? *ULTRA* doesn't offer answers – it gives provocations.
Accessible, nontinear, and laced with dark imagination and rigorous speculation, *ULTRA* is essential reading for anyone intrigued or terrified by what lies beyond the algorithmic event horizon.

University of Groningen Press

ISBN 978-94-034-3132-1